Global Biogeography of the Soil Virosphere

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

[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

[6]Department of Agronomy, Iowa State University, Ames, IA USA

*Correspondence: emily.graham@pnnl.gov

ABSTRACT

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 (GSV) Atlas, a comprehensive dataset from 2,953 soils that comprises 616,935 uncultivated virus genomes (UViGs) and 38,508 unique viral operational taxonomic units (vOTUs). Rarefaction curves from the GSV Atlas indicate that most soil viral diversity remains unexplored, further underscored by high spatial turnover and low rates of shared vOTUs 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 the most extensive characterization of soil viral diversity to date 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 considerable 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[1–4]. Early metagenomic investigations suggest that soil viruses are more genetically diverse than nearly any other organism and may significantly impact global biogeochemistry[2,3,5,6]. Still, less than 1% of publicly available viral metagenomic sequences are from soil[7], reflecting the lack of knowledge about soil viruses and their ecological roles[5,8,9].

High soil viral diversity may be due to the structural, nutrient, and/or physico-chemical heterogeneity of soil compared to other ecosystems[2,10–13], as well as the high diversity of their microbial hosts. Indeed, soil viral abundance and composition vary with factors such as pH, habitat, moisture, land use, carbon and nutrient content, and temperature[13–17]. While less is known about soil viral activity, close to 60% of the soil virosphere may be involved in active infections[18], consistent with high activity observed in marine and other systems[5,9,19–23].

Whether common ecological patterns observed in microorganisms apply to the soil virosphere remains an open question; initial studies of the soil virosphere indicate that ecological processes in viral and microbial communities are at least partially decoupled[10,13,24]. A major finding is that soil viral community turnover may occur over shorter spatial and temporal scales than microbial communities[10,13,24]. For instance, spatial viral turnover has been shown to be over 5 times higher than microbial community turnover in agricultural soils across an 18m transect[10], and only 4% of peatland ‘viral operational taxonomic units' (vOTUs) are shared across continents[25]. Other studies note the possibility for long-distance soil viral dispersal through atmospheric[26] or aquatic transport[27] consistent with low turnover. These contrasting results

55 indicate a lack of consensus surrounding the spatial patterns of soil viruses and the need for
56 large-scale surveys of the soil virosphere.

57 Importantly, soil viruses can influence biogeochemical cycling, antibiotic resistance, and
58 other critical soil functions[28–33] by releasing carbon and nutrients during host infection or by
59 altering host metabolism via auxiliary metabolic genes (AMGs)[11,18,22,34–42]. While the
60 characterization of viral AMGs in soil is nascent[43], marine systems demonstrate the breadth of
61 functions ripe for potential discovery in soil[38]. More than 200 viral AMGs encoding functions
62 related to carbon, nitrogen, sulfur, and phosphorus cycling; sporulation and stress tolerance; and
63 toxin resistance have been detected in marine systems[38]. In contrast, only a handful of these
64 functions have been identified as soil viral AMGs[15,17,18,27,43–46]. AMGs encoding carbohydrate
65 metabolism genes in particular may be present in soils, including a few that have been
66 experimentally validated[11,13,18,45–48].

67 Accordingly, understanding the role of viruses in soil ecosystems is one of the most
68 pressing current challenges in microbial ecology[49]. Despite the expansion of studies
69 characterizing soil viral diversity, activity, and impact[5,9,17,45,46], a comprehensive description of
70 the global soil virosphere has yet to be performed. Such a description is necessary to begin to
71 address questions regarding the spatiotemporal dynamics, physicochemical interactions, host
72 organisms, and food web implications of the soil virosphere. We present a meticulous
73 compilation and curation of the Global Soil Virus (GSV) Atlas. This atlas represents the most
74 extensive collection of viruses in soil metagenomes to date, encompassing contributions from
75 prominent repositories, ecological networks, and individual collaborators.

76

77 **RESULTS**

78 *Global Soil Virus (GSV) Atlas.*

79 In total, we amassed 1.25×10^{12} of assembled base pairs (bp) across 2,953 soil samples,
80 including 1,552 samples that were not previously available in the United States Department of
81 Energy's (DOE), Joint Genome Institute's (JGI) IMG/M database (Figures 1 and 2). These
82 samples were then screened for viruses, yielding 616,935 uncultivated virus genomes (UViGs) of
83 which 52,211 were of sufficiently high quality for further investigation (see methods). Newly
84 contributed sequences clustered into 3,613 vOTUs of which 317 clustered with existing viral
85 sequences in IMG/VR. The vast majority associations with IMG/VR were with sequences
86 previously uncovered from soil habitats (Figure 2b).

87 We also collected an array of associated metadata including environmental parameters
88 describing each sampling location from the SoilGrids250m database[50]. We assayed a wide
89 variety of soil environments that ranged from bulk density of 0.24-1.56 (bulk density; cg/cm^3),
90 cation exchange capacity of 6.8-71 (CEC; mmol(c)/kg), nitrogen content of 0.19-22.4 (N; cg/kg),
91 pH ranging from 4.3-8.5, soil organic C of 1.9-510.9 (SOC; dg/kg), and clay content of 2.7-57.1
92 (g/kg) (Figure 1).

93 UViGs clustered into 38,508 vOTUs at the species-like-level[51], of which 3,296 were
94 previously unrepresented in IMG/VR (Figure 2a-b). Only 13.9% of the GSV Atlas' vOTUs
95 appeared in more than one sample, and less than 1% were present in more than 5 samples. At
96 higher taxonomic levels, we found 21,160 clusters at the genus level, and 7,598 clusters at the
97 family level[52]. This equates to an average of 40.01, 35.48, and 24.91 unique viral clusters per
98 sample at the species, genus, and family levels. 38,278 out of 38,508 vOTUs (99.4%) had at least
99 one member that was assigned to a taxon by geNomad.

100 We identified 1,573,689 viral genes, of which only 66,684 (4.2%) were found in at least
101 one annotation database (1,022, 704, and 252 unique KO, Pfam, and CAZy annotations,
102 respectively, Figure 2). After filtering to putative AMGs[46], we found 5,043 genes that mapped to
103 83 KEGG pathways (1,941 KO and 3,357 CAZyme, some genes were associated with both KO
104 and CAZyme). The mean putative AMG abundance (gene copies per sample) was 18.99 genes
105 (median = 4, suggesting this distribution was skewed by a few samples with high total putative
106 AMG abundance). The KEGG pathways with the most mapped putative AMGs were Galactose
107 metabolism, Glycosaminoglycan degradation, Starch and sucrose metabolism, N-Glycan
108 biosynthesis, and Folate biosynthesis; while the most abundant gene annotations associated with
109 AMGs were GlycosylTransferase Family 4 (GT4), Carbohydrate-Binding Module Family 50
110 (CMB50), 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase (K01737), Glycoside
111 Hydrolase Family 73 (GH73), and 7-carboxy-7-deazaguanine synthase (K10026) (Figure 2c-d).
112 In contrast to the saturation observed in rarefaction curves for the taxonomy of microbial
113 sequences and microbial genes annotated by Pfam, rarefaction curves of soil viruses for
114 individual samples (vOTUs and viral clusters at the genus and family levels) and their genes
115 (annotated by Pfams) did not reach an asymptote (Figure 3). The number of new and unique
116 vOTUs and viral clusters at the family level (Figure 3b and c) was linearly related to sequencing
117 depth, while viral Pfams displayed slight curvature. These linear relationships were observed
118 when considering 2TB of metagenomic sequencing—4-fold more sequencing depth than any
119 other soil metagenome in this study and 40-fold more than the Joint Genome Institute (JGI)
120 recommended sequencing depth for soil samples (45 GB). When considering cumulative unique
121 attributes versus sequencing depth (Figure S1), relationships in vOTUs and viral clusters at the
122 family level displayed slight curvature, while viral Pfams neared saturation.

123

124 *Microbial Hosts of Soil Viruses.*

125 We connected a total of 1,450 viruses to putative hosts of 82 different bacterial and
126 archaeal orders with CRISPR spacers. This equates to of 2.78% viral sequences passing QA/QC
127 associated with CRISPR spacer hits; roughly 70% more host assignments than in another recent
128 assessment[5,53]. While we observed a maximum of 73 vOTUs per host (i.e., CRISPR spacer), the
129 mean overall vOTU per host ratio was 0.42 (median = 0), reflecting the predominance of unique
130 host associations for individual vOTUs.

131 Out of 1223 samples with at least one vOTUs assigned to a host, only 72 samples had an
132 average of more than one host sequence per vOTU, underscoring the low abundance of detected
133 hosts across all soils. An average of 0.64 unique host orders were detected per sample, with a
134 maximum ratio of CRISPR spacer hits to viral sequences of 73. Further, samples with a high
135 ratio of vOTU:host almost exclusively were matched to host sequences from a single microbial
136 order, reflecting high phylogenetic conservation of host associations. Of the 10 samples with the
137 highest CRISPR spacer sequence to viral sequence ratio, only 1 contained a CRISPR spacer
138 matching more than one bacterial or archaeal order, and it only contained matches to two orders.

139 We found that the most abundant host taxa were distributed across distantly-related
140 phyla, including members of prominent soil orders such as Pseudomonadales, Burkholderiales,
141 Acidobacteriales, and Bacteroidales (Figure 4b). The abundances of host organisms belonging to
142 Acidobacteriales, Oscillospirales, Pedosphaerales, and Geobacterales were positively associated
143 with soil nitrogen, SOC, and CEC; while host members of Enterobacterales, Obscuribacterales,
144 Mycobacteriales, Pseudomonadales, and Streptomycetales were positively associated with soil
145 bulk density, and to a lesser extent, pH and clay content.

146

147 *Metabolic potential encoded by the soil virosphere.*

148 Because functional annotations of viral genes were distributed across an array of
149 functions, we screened all viral genes (regardless of assignment by the AMG detection pipeline)
150 against KEGG pathways to understand relationships among genes in the context of known
151 metabolic processes. Across the entire soil virosphere, we uncovered many portions of KEGG
152 pathways that were mostly complete, including functions involved in amino acid and sugar
153 metabolism and biosynthesis, antibacterial mechanisms, nucleotide synthesis, and other viral
154 functions (e.g., infection strategies) (Supplemental Material). Of note, we found nearly complete
155 portions of energy-generating pathways including the pentose phosphate pathway and F-type
156 ATPase-mediated portions of photosynthesis. We also observed a prevalence of
157 lipopolysaccharide (LPS) pathway-related genes that may be important as host receptors for
158 bacteriophage and prevention of superinfection[54–56].

159

160 **DISCUSSION**

161 *Global Soil Virus (GSV) Atlas.*

162 The GSV Atlas demonstrates the immense, unexplored taxonomic and functional
163 diversity of the soil virosphere. Soil viral diversity in the GSV Altas appeared to be largely
164 distinct from other global habitats. Nearly 80% of GVS Atlas sequences that clustered with
165 existing sequences in IMG/VR were attributed to soil or soil-like habitats (i.e. 'other terrestrial'
166 or 'plant-associated'/rhizosphere), underscoring the unique composition of the soil virosphere
167 relative to more well-studied marine and human environments. Additionally, we found low
168 numbers of shared vOTUs and viral clusters across samples indicative of high spatial turnover

169 (i.e., changes in soil viral community composition across spatially-distributed samples). Recent
170 studies have estimated soil viral diversity is high both relative to other viral habitats and relative
171 to soil microbial diversity[8,10,12,13,27,44,57]. However, these estimates have been limited by the sheer
172 amount of viral and microbial 'dark matter' for which no functional or taxonomic assignment is
173 known[34,43,49,57,58]. Towards this end, the diversity encompassed by the GSV Atlas can serve as a
174 community resource for potentially characterizing this unknown fraction of the soil virosphere.

175 Analysis of the GSV Atlas suggests that extreme spatial variability (compared to
176 microorganisms and macroorganisms) is a key feature of the soil virosphere at the global scale.
177 While rapid viral spatial turnover was recently observed across short spatial scales (<10-
178 20m)[10,13,59], there has been no such demonstration of viral biogeography across the world. We
179 propose that high rates of spatial turnover could result from low dispersal rates or distinct
180 temporal dynamics of viral communities relative to other microorganisms. For example, while
181 dormant microorganisms and relic DNA can persist for months or more[35,60–63], the burst of viral
182 cells associated with active infections may generate short-lived pulses of distinct viral
183 communities that do not contribute to relic DNA.

184 Together, these factors make characterizing the soil virosphere an immense challenge for
185 the coming decade. When examining individual soil samples, the number of new and unique
186 viral attributes (e.g., vOTUs, clusters, Pfams) observed in this study is linearly related to
187 sequencing depth, suggesting that the discovery of new soil viruses is likely to continue with
188 increasingly deep sequencing (Figure 3). This contrasts with rarefaction curves of the soil
189 microbiome and of microbial hosts of soil viruses, which both tend to asymptote well before
190 typical sequencing depths employed in soil microbial investigations. Still, when looking at the
191 cumulative number of unique viral attributes detected in all samples collectively (Figure S1),

192 many viral attributes begin to saturate with sequencing depth. This finding suggests that while
193 individual samples do not capture soil viral diversity, we can begin to constrain the extent of
194 diversity when sequences from thousands of existing samples are aggregated.

195 Assembling the GSV Atlas enables us to begin to describe the enormous functional
196 diversity in the soil virosphere. After filtering viral genes to putative AMGs (see methods),
197 KEGG pathways among those represented by the highest number of putative AMGs were
198 associated with a suite of major soil carbon cycling processes (KO:00052, Galactose
199 metabolism; KO:00500, Starch and sucrose metabolism). Likewise, at the level of gene
200 annotations, the most common putative AMGs suggested a role for viruses in soil carbon cycles.
201 Glycosyltransferases 4 (GT4) and glycosylhydrolases 73 (GH73) encode enzymes involved in
202 the metabolism and/or production of common sugars found in soils including sucrose, mannose,
203 glucosamine, and maltose[64,65]; and Carbohydrate Binding Module 50 (CBM50) has been
204 previously linked to chitin – one of the most abundant carbon molecules in soil[66].

205 Collectively, these genes are involved in the decomposition of chemically complex
206 organic material into substrates for microbial energy generation and growth. Our results are
207 consistent with previous works from single locations which have hinted at a wide range of
208 possible soil viral AMGs, including glycoside hydrolases, carbohydrate esterases, and
209 carbohydrate-binding modules[18,34,47]. Given that a large proportion of soil microorganisms are
210 infected by viruses at any given time[67], AMGs encoded by soil viruses have the potential to
211 impact global biogeochemical cycles[18,27,34,47,48]. The thousands of putative AMGs identified here
212 represent the most extensive survey to date and further impress the importance of the soil
213 virosphere as a reservoir for biogeochemical potential.
214

215 *Microbial Hosts of Soil Viruses.*

216 Unraveling relationships between viruses and their host communities is imperative to
217 understanding the impact of the soil virosphere on soil processes. Host presence should be tightly
218 coupled to viral abundance, and in turn, these linkages can be mediated by spatial, temporal, and
219 environmental factors[18,68,69,70]. These linkages are also dependent on viral host range (i.e., host
220 specificity) — higher host specificity should lead to a stronger coupling between microbial and
221 viral abundance and community composition. Viral host specificity is also associated with
222 ecological factors that impact microbial community composition and may result in tradeoffs
223 between viral growth rates versus the breadth of the host range[16,71–77]. These dynamics are
224 unresolved in soils, leading to open questions regarding associations between viral host
225 specificity and microbial community functions.

226 Our analysis of the GSV Atlas indicates that a limited host infection range may be a
227 dominant trait across soil viruses. Across the GSV Atlas, there were few hosts per vOTU on
228 average (mean = 0.42 hosts/vOTU), and of vOTUs associated with multiple host sequences, the
229 vast majority were linked to multiple hosts of the same phylogenetic clade. While high host
230 specificity has historically been the prevailing paradigm, our work contrasts recent studies
231 suggesting that some soil viruses may have broader host ranges than viruses in other habitats[78–
232 80].

233 As a whole, we identified taxonomically diverse soil microorganisms as viral hosts
234 (Figure 4). Host sequences spanned nearly every major soil microbial clade, supporting a suite of
235 recent studies that detected soil viral hosts classified as Acidobacteria, Verrucomicrobia,
236 Deltaproteobacteria, and Gammaproteobacteria[27,34,47]. The ultimate impact of viral predation on
237 soil functions is at least partially associated with the taxonomic distribution of hosts—viral

238 impacts may be larger when their hosts contribute to key ecosystem functions. The taxonomic
239 breadth of the microbial hosts of soil viruses detected here suggests a role for the soil virosphere
240 in nearly every soil habitat on Earth, as the microbial hosts found here occupy a wide range of
241 ecological niches. Moreover, some hosts were susceptible to changes in the environment,
242 underscoring the role that soil heterogeneity may play both in viral diversity (as described above)
243 and in shaping viral-host dynamics.

244 Relationships between environmental factors and host taxa may reflect environmental
245 filtering on host communities (which in turn, determine the amount and type of viruses present)
246 or on viruses directly which subsequently impacts host community composition[18,28,30,81,82]. Viral
247 infections have been previously linked to soil parameters including moisture[17,46,83] and carbon
248 and nitrogen content[11]. In our analysis, bulk density may serve as a proxy for the role of
249 hydrologic connectivity in structuring the soil virosphere and its hosts. For example,
250 heterogeneity in the soil matrix (e.g., 'spatial refuges' created by low hydrologic connectivity) is
251 thought to impact viral activities[10,84]. Nutrient amendments are also considered to be drivers of
252 the soil virosphere[10,84], supporting the relationships we observed between carbon, nitrogen, and
253 host taxa.

254

255 *Metabolic Potential Encoded by the Soil Virosphere.*

256 We detected many hallmarks of viral activity in the soil virosphere, including genes
257 associated with cell wall lysis, DNA repair/replication, and other signatures of infection. For
258 instance, across the entire soil virosphere, genes associated with DNA mismatch repair
259 (KO:03430), homologous recombination (KO:03440), and base excision repair (KO:03410) were
260 prevalent (Figure S2). Folate biosynthesis was also commonly encoded by the soil virosphere.

261 Folate and other B vitamins are key to bacterial growth (KO:00670, KO:00790, Figure S3[86]) and
262 therefore may be logical targets for soil viruses[85]. Bacterial secretion systems (KO:03070, Figure
263 S4), which may be evolutionarily derived from phages[87–89], were also found in many soil virus
264 genomes. Type IV secretion systems in particular can be used by bacteria to secrete toxins[90] or as
265 a method for DNA transfer through cell membranes[91]. Finally, genes involved in the
266 *Caulobacter* cell cycle (KO:04112, Figure S5), which has a distinct pattern of cell division[92]
267 were rife across soils and may be promising indicators of viral infections. The prevalence of viral
268 genes associated with central microbial functions highlights the potential importance of viral
269 activity in soil habitats and the need for targeted approaches to quantify the extent and impact of
270 viral gene expression.

271 The GSV Atlas also contains many viral amino acid biosynthesis and degradation
272 pathways that could be critical in viral infection and replication (e.g., KO:00250, KO:00260,
273 KO:00270, KO:00330, KO:00340, Figure S6). Amino acids are building blocks for new cellular
274 material and can support soil biogeochemical cycles, as they serve to enhance the carbon cycle
275 through priming effects and/or enhanced nutrient availability[93–96] (e.g., KO:00500, KO:00052,
276 KO:00051, Figure S7). Collectively, these pathways demonstrate several possibilities for critical
277 impacts of soil viruses on processes that are central to microbial metabolism and biogeochemical
278 cycling of elements in soil.

279 Beyond these pathways, we highlight three KEGG pathways with near-complete portions
280 represented in the GVS Atlas: F-type ATPase (KO:00190), pentose phosphate pathway
281 (KO:00030), and LPS (KO:00540). Five of seven subcomponents of the F-type ATPase were
282 detected in the soil virosphere. Given the evolutionary similarities between V- and F-type
283 ATPase in particular[97], the lack of any V- or A-type ATPase components is notable in light of

284 the near complete F-type ATPase. Though there is some basis for F-type ATPases to assist in
285 viral replication[98–100], we also note the possible involvement of F-type ATPase in photosynthetic
286 energy generation[101]. Given the prominence of photosynthetic AMGs in marine systems[102–104],
287 we highlight the possibility of a viral F-type ATPase as a soil AMG. The pentose phosphate
288 pathway is also a prevalent and important AMG found in marine ecosystems, where viral
289 infection diverts carbon away from the Calvin cycle towards the pentose phosphate pathway and
290 allows for an 'express route' of energy generation at the expense of host carbon metabolism
291 (reviewed in [102]). Finally, we observed nearly complete LPS pathways in the GSV Atlas. Phages
292 often carry depolymerases and other enzymes that target LPS or similar outer membrane
293 components to facilitate binding and entry[54,55]. However, the representation of members of the
294 LPS synthesis pathway in identified AMGs indicates that phage may be also working to change
295 the function of the pathway post-infection, potentially to prevent superinfection as has been
296 observed in specific phages[56]. Collectively, we propose that F-type ATPase, pentose phosphate
297 pathway, and LPS may be interesting targets for more targeted investigations into the role of the
298 virosphere in soil health and microbiome functions.

299

300 *Future Recommendations.*

301 The field of soil viral ecology is poised for rapid expansion; yet there are several
302 challenges to fully characterizing soil viral diversity and function. Overcoming these
303 methodological and ecological hurdles will require broad participation from global researchers.
304 Below, we present a summary of issues, from our perspective, facing the current generation of
305 soil viral ecologists along with suggestions for surmounting them (Box 1).

306 First, we propose investments in methodological advances to improve viral detection and

307 resolve genomic 'dark matter'. Metagenomic sequencing enables the detection of thousands of
308 viruses per soil sample, yet the number of viruses detected in soil metagenomes has remained
309 relatively flat over time[5]. In part, this is because soil metagenomics from shotgun sequencing is
310 highly fragmented, leading to lower quality UViGs[105–107]. Identifying novel viral sequences and
311 assigning viruses to microbial host organisms are also limited by the extent of our knowledge of
312 viral diversity; thus expansion of the known virosphere is needed. Promising new methods
313 include experimental verification of viral activity[45,108], size fractionation ('viromics')[8,10,18,57],
314 viral isolation[109,110], optimized viral nucleic acid extraction[111], microscopy[45], combined
315 metagenomic assembly[5], and long-read and/or single-cell sequencing[112–114].

316 Knowledge about soil viral diversity and function is also limited by gaps in field and
317 laboratory experiments. The GSV Atlas demonstrates that extensive, spatially-explicit sampling
318 is needed to capture the high spatial turnover of the soil virosphere. In particular, the spatial
319 coverage of most 'global' ecological studies, including this one, often suffers from large data
320 gaps[115]. Concerted efforts are needed to sample wide spatial domains, including historically
321 undersampled regions, given the high viral diversity uncovered by the GSV Atlas. Although we
322 did not assess temporal dynamics, temporally-explicit approaches are likewise needed to
323 characterize temporal dynamics in soil viral communities. Further, our functional annotation of
324 viral contigs revealed diverse genes associated with functions relevant to both viral and
325 microbial communities, and it is impossible to know the true functions of viral genes without
326 targeted functional assays. We therefore propose that experiments targeting the expression and
327 auxiliary metabolic function of viral genes are needed to properly assess AMGs in viral
328 communities.

329 Finally, we still know relatively little about the ecological drivers of soil virus
330 distribution and how to represent these mechanisms in process-based models. Extreme soil
331 virosphere diversity renders some common microbial ecology statistical methods unfeasible,
332 including those often used to test ecological principles (e.g., ordinations, distance-decay,
333 richness, etc). This highlights the need for innovative statistical approaches to interpret the soil
334 virosphere. While process-based models have made tremendous improvements in representing
335 soil carbon cycles, the soil virosphere is a key missing component. New ecological theory may
336 be needed to enable the inclusion of the soil virosphere in biogeochemical models.
337
338 *Conclusions.*
339 The GSV Atlas is a new public resource that can help generate hypotheses and provide
340 insight into some of the pressing challenges of viral ecology. We uncovered 616,935
341 uncultivated virus genomes from global soil samples to show the extreme diversity, spatial
342 turnover, and functional potential of the soil virosphere. This includes a wide taxonomic array of
343 microbial hosts of soil viruses, key functions associated with soil carbon cycles, and an
344 assortment of viral metabolisms that may be critical to deciphering viral ecological principles in
345 the soil ecosystem. We specifically highlight F-type ATPase, the pentose phosphate pathway,
346 and LPS-related genes, as well as enzymes involved in carbohydrate metabolism, as fruitful
347 areas for further investigation. Our work scratches the surface of the soil virosphere and serves as
348 a basis for tool, theory, and model development to further advance soil ecology,
349 biogeochemistry, ecology, and evolution.
350
351 **Data availability.** All data will be publicly available upon acceptance. During review, please

352 refer to the link for reviewers provided to the editor.

353

354 **Acknowledgements.** This work was supported by the U.S. Department of Energy (DOE), Office
355 of Biological and Environmental Research (BER) as part of BER's Genomic Sciences Program
356 (GSP) under FWP 70880.

357

358 **Author Contributions.** EBG, RW, JKJ, DPE, NCK, and JEM conceived of this project. EBG,
359 RW, APC, RYN, MN, and AC conducted data analysis. EBG, RW, APC, RYB, NCK, JKJ,
360 KSH, and JEM contributed to manuscript drafting and revisions. The Soil Virosphere
361 Consortium contributed metagenomic data and reviewed the manuscript.

362

363 **Competing interests.** The authors declare no competing interests.

364

FIGURES

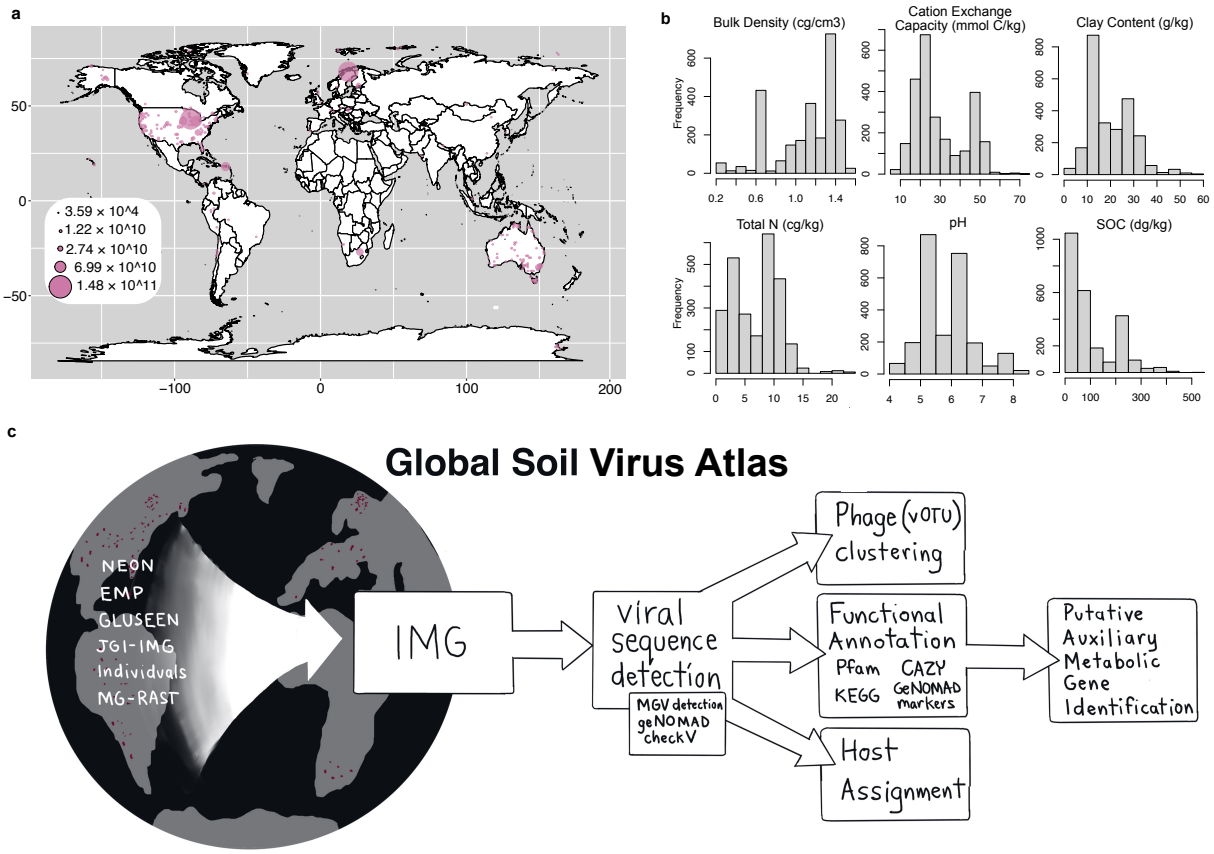


Figure 1. Data collection and workflow. a, Global distribution of samples, scaled by assembled base pairs. In order horizontally, histograms of (**b**) mean soil bulk density (cg/cm^3), cation exchange capacity (mmol(c)/kg), clay content (g/kg), total nitrogen content (cg/kg), pH, and soil organic carbon (dg/kg) associated with our samples from the SoilGrids250 database (0-5 cm). **c,** Sequence processing pipeline.

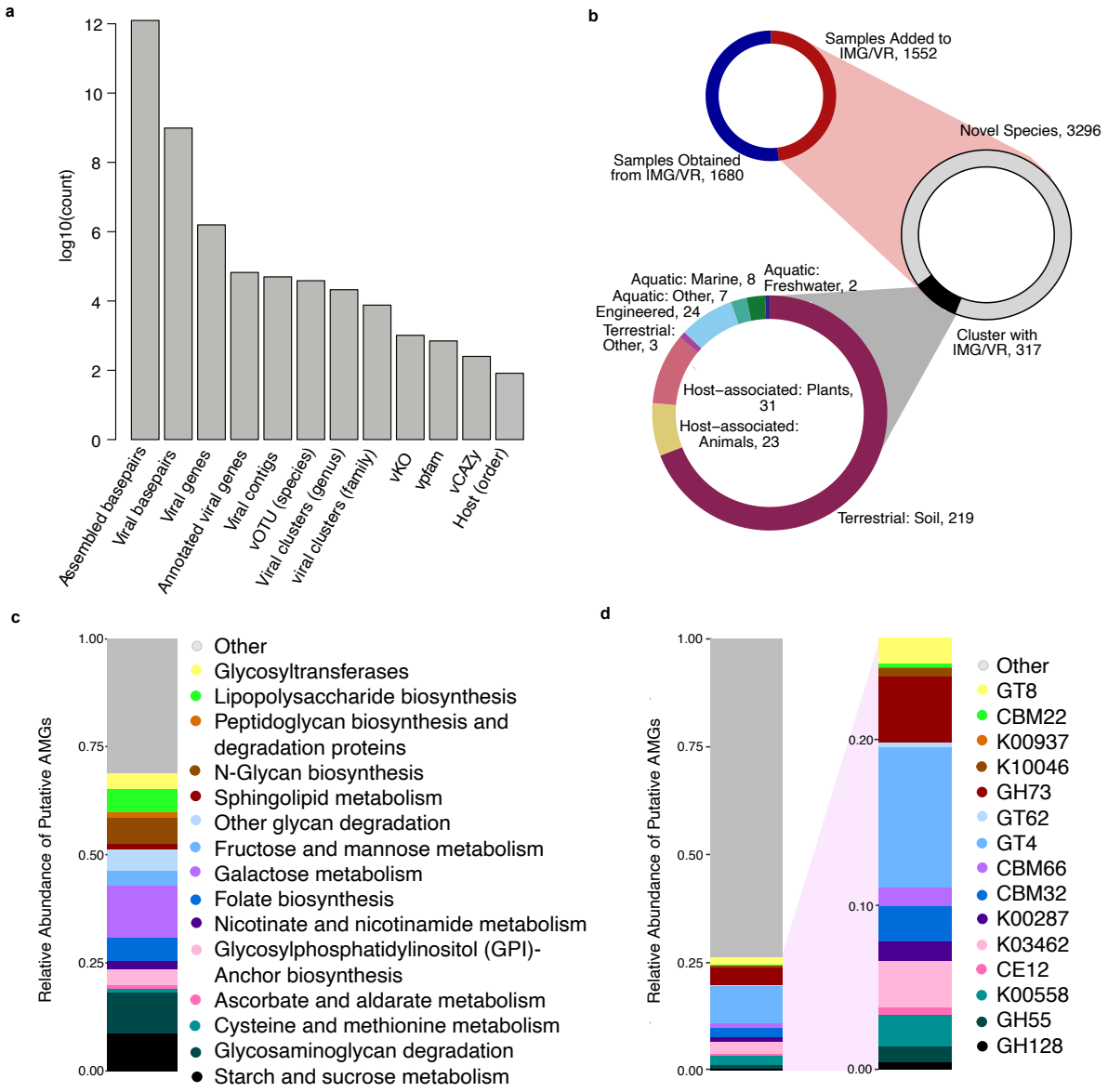
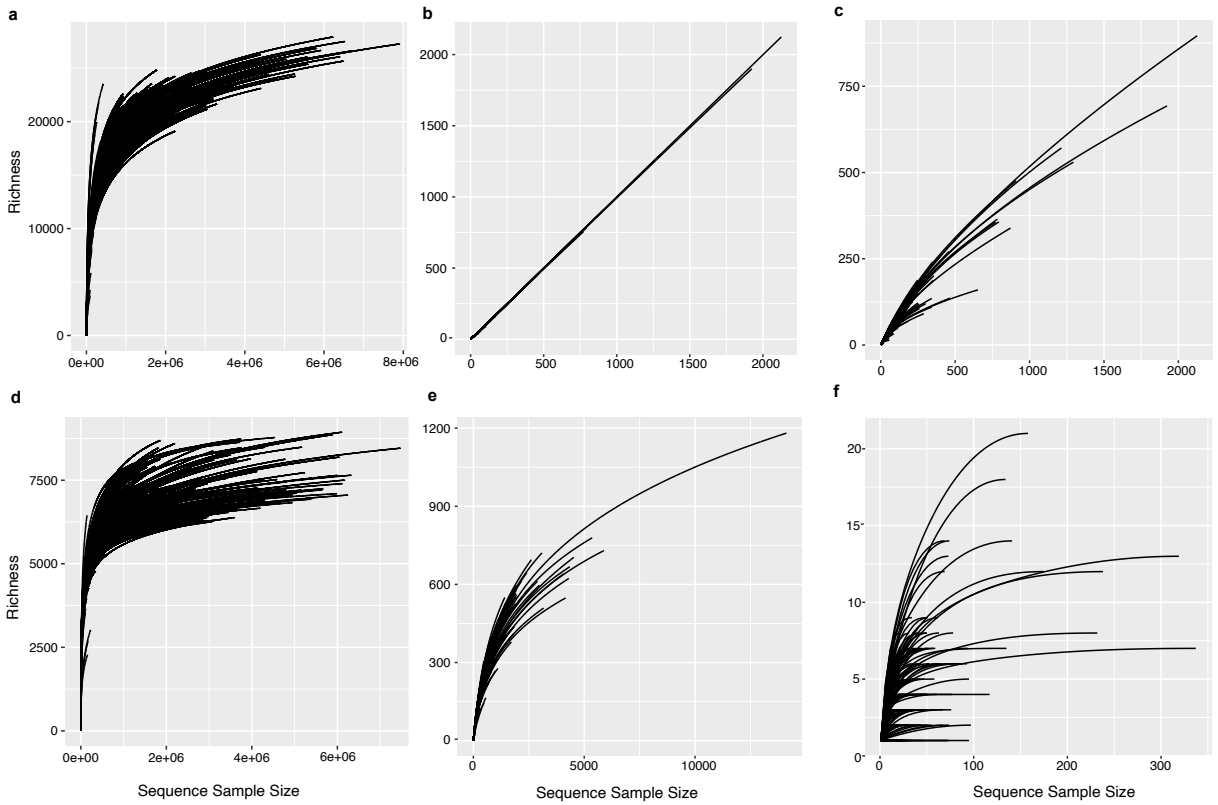


Figure 2. Data description. a, shows the count of each category across the full dataset. **b**, shows the proportion of samples obtained from IMG/VR versus the number of new samples contributed (upper left). Within new samples, we identified 3,613 vOTUs of which 317 clustered with sequences already in IMG/VR (middle). Sequences in IMG/VR that clustered with vOTUs containing new sequences were mostly associated with soil habitats (lower left). **c-d,** relative abundance of putative AMGs grouped by KEGG pathway **(c)** and by annotation **(d)**.



379

380 **Figure 3. Rarefaction curves. a,** taxonomy from whole metagenomic sequences, **b**, vOTUs at

381 the species level, **c**, viral clusters at the family level, **d,** Pfams from whole metagenomic

382 sequences, **e**, Pfams from UViGs, **f**, CRISPR spacer-based host assignment of uViGs.

383

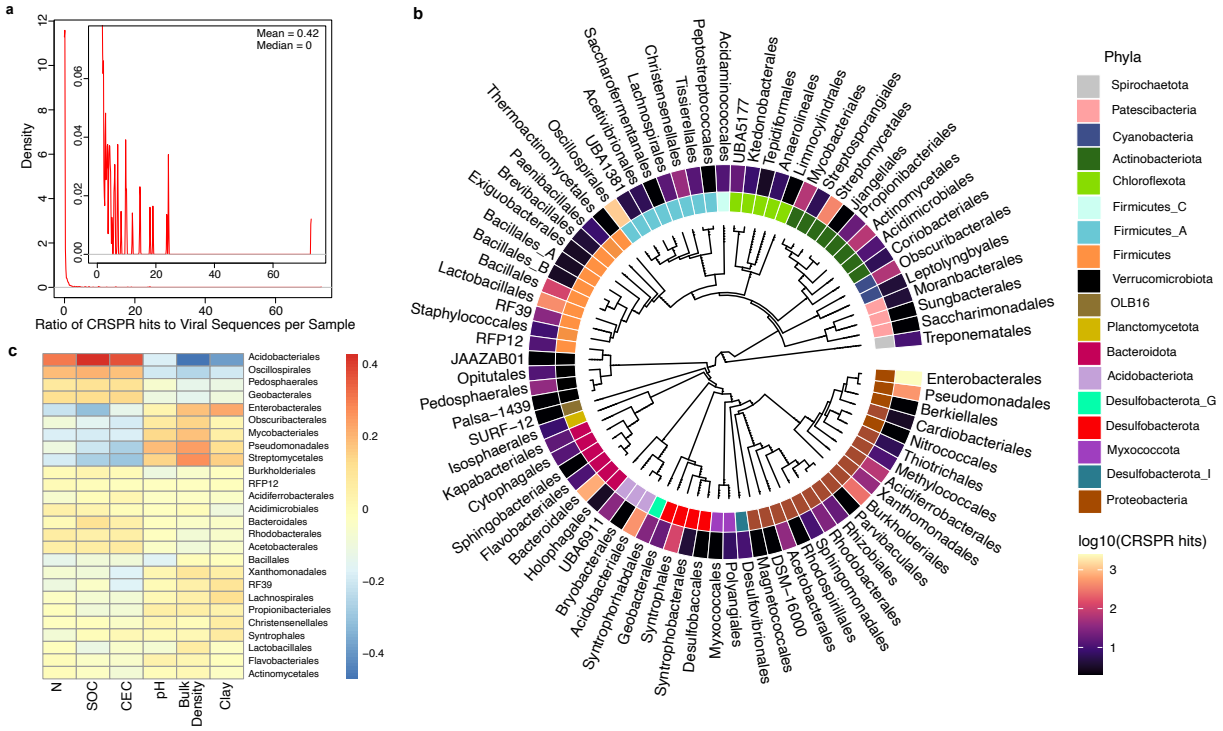


Figure 4. Relationships between soil viruses and their hosts. a) Cumulative Distribution Function plot showing the ratio of CRISPR spacer hits to viral sequences per sample. Inset shows a zoomed portion of the plot from 0 to 0.08 along the y-axis. **b)** Phylogenetic tree of bacterial hosts at the order level. Phylum 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**) Correlation between host abundance (defined as total CRISPR spacer hits per microbial order) and environmental parameters from SoilGrids250m. Color denotes Spearman rho. For inclusion in the heatmap, host order must be present in more than 5 samples.

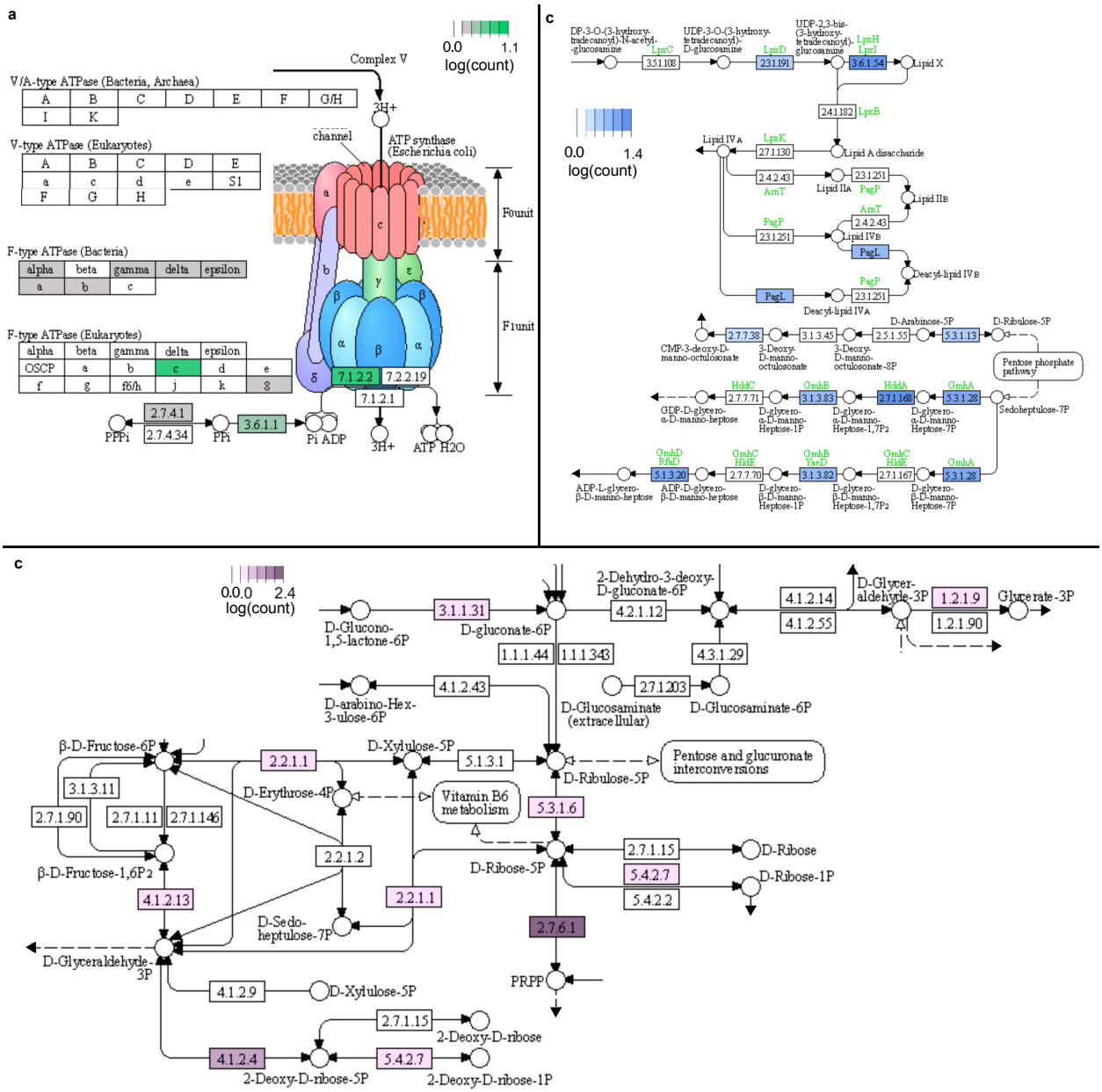


Figure 5. Metabolic potential encoded by the soil virosphere. a, F-type ATPase (KO:00190), **b**, lipopolysaccharide pathway (KO:00540), and **c**, pentose phosphate pathway (KO:00030). KEGG pathways are cropped and/or simplified to enhance visualization. Graphics are rendered by Pathview. Color scales denote the total of the total abundance across the entire soil virosphere.

401 **Box 1. Recommendations for Uncovering Soil Virosphere Diversity and Function.**

402 ● Technical advances to improve soil virus identification from shotgun metagenomics, long
403 read sequencing, and/or broader adoption of filtered/targeted sequencing approaches
404 ● Expansion of the known virosphere to facilitate tool development
405 ● Improved methods, both experimental and computational, to link soil viruses to their host
406 organisms.
407 ● Spatially and temporally-explicit sampling is needed to capture the high spatial turnover
408 of the soil virosphere.
409 ● Experiments to capture expressed auxiliary metabolic functions of viral genes
410 ● Studies and improved metadata to understand the ecological drivers of soil virus
411 distributions
412 ● Novel statistical approaches for ecological interpretation of sparse data
413 ● New modeling representations of viral impacts on soil carbon and nutrient cycling

414

415 **METHODS**

416 *Data Collection and Curation.*

417 We collected a total of 2,953 soil metagenomic samples from soils from major
418 repositories and ecological networks including the JGI Integrated Microbial Genomes and
419 Microbiomes (IMG/M) platform, MG-RAST metagenomics analysis server, Global Urban Soil
420 Ecological Education Network (GLUSEEN), Earth Microbiome Project (EMP), and National
421 Ecological Observatory Network (NEON) plus submissions from individual collaborators. This
422 included 1,552 samples not previously included in IMG/M (Figure 1 and 2).

423 For samples collected via JGI IMG/M, we retrieved all studies with GOLD[116] ecosystem
424 type of "Soil" as of August 2020. We manually curated metagenomic sequences to remove
425 misclassified data as follows. We removed samples from studies with the following: (1) GOLD
426 ecosystem types: Rock-dwelling, Deep subsurface, Plant litter, Geologic, Oil reservoir, Volcanic,
427 and Contaminated; (2) GOLD ecosystem subtypes: Wetlands, Aquifer, Tar, Sediment, Fracking
428 Water, and Soil crust; (4) words in title: wetland, sediment, acid mine, cave wall surface, mine
429 tailings, rock biofilm, beach sand, Petroleum, Stalagmite, Subsurface hydrocarbon microbial
430 communities, Vadose zone, mud volcano, Fumarolic, enriched, Composted filter cake, Ice
431 psychrophilic, oil sands, groundwater, Contaminated, rock biofilm, Deep mine, coal mine fire,
432 Hydrocarbon resource environments, Marine, enrichment, groundwater, mangrove, saline desert,
433 Hydroxyproline, Rifle, coastal, compost, biocrust, crust, Creosote, soil warming, Testing DNA
434 extraction, and/or Agave; (5) GOLD geographic location of wetland; and (6) GOLD project type
435 of Metagenome - Cell Enrichment. Additionally, sample names that indicated experimental
436 manipulation (e.g., CO_2 enrichment or nitrogen fertilization) or were located in permafrost layers

437 were manually excluded. This resulted in 1,480 curated metagenomes from publicly available
438 data in IMG/M.

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

442 In parallel, we retrieved mean values for soil parameters from the SoilGrids250m
443 database from 0-5 cm[50]. SoilGrids250m is a spatial interpolation of global soil parameters using
444 ~150,000 soil samples and 158 remote sensing-based products. Here, we focus on six parameters
445 often associated with soil microbial communities: bulk density, cation exchange capacity,
446 nitrogen, pH, soil organic carbon, and clay content. Because our focus on spatial dynamics and
447 soils were collected at various times, we did not include temporally dynamic variables such as
448 soil moisture or temperature in our set of environmental parameters, though we acknowledge
449 they may have profound impacts on the soil virosphere.

450

451 *Assembly and Annotation of Samples added to IMG/M.*

452 To standardize data analysis across all samples, the 1,552 soil metagenomic samples not
453 collected from IMG/M were analyzed using the JGI's Metagenome Workflow[117]. In brief,
454 samples were individually assembled using MetaSpades v. 3.1. 1,476 of the 1,552 assembled soil
455 samples passed default quality control thresholds[118], yielding 133 gigabases of assembled DNA
456 in 241,465,924 contigs. Additionally, three very large metagenomes (>1TB each) were
457 assembled separately due to computational limitations in standard workflows[119]. The resulting
458 assemblies were assigned GOLD identification numbers and imported into IMG/M and

459 processed using version 5.0.0 of the IMG/M Metagenome Annotation Pipeline to align with data
460 obtained directly from IMG/M[117].
461
462 *Virus identification, clustering, and host prediction.*
463 We performed an initial identification of viral contigs using using a modified version of
464 the IMG/VR v3's virus identification pipeline (code available at
465 https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline)[52,53]. The pipeline
466 identifies viruses based on the presence of 23,841 virus protein families, 16,260 protein families
467 of microbial origin from the Pfam database[120] and VirFinder[121] to identify putative virus
468 genomes in contigs that were at least 1 kb long. During the course of this study, geNomad
469 (version 1.3.3)[122], a tool for virus identification with improved classification performance was
470 released and incorporated into our pipeline to improve prediction confidence and perform
471 taxonomic assignment. We further processed predicted viral sequences using CheckV v1.0.1
472 (database version 1.5)[123] to assess the quality of the virus genomes. As this study focused on
473 non-integrated virus genomes, contigs that were flagged by either geNomad or CheckV as
474 proviral were discarded. From the remaining contigs, virus genomes were selected using the
475 following rules: (1) contigs of at least 1kb with high similarity to genomes in the CheckV
476 database (that is, that had high- or medium-quality completeness estimates) or that contained
477 direct terminal repeats were automatically selected; (2) contigs longer than 10kb were required to
478 have a geNomad virus score higher than 0.8 and to either encode one virus hallmark (e.g.,
479 terminase, capsid proteins, portal protein, etc.) .), as determined by geNomad, or to have a
480 geNomad virus marker of at least 5.0; (3) contigs shorter than 10 kb and longer than 5 kb were
481 required to have a geNomad virus score higher than 0.9, to encode at least one virus hallmark,

482 and to have a virus marker enrichment higher than 2.0. This resulted in 49,649 viral contigs that
483 we used for downstream analysis. All viral contigs are available at: [refer to link provided for
484 reviewers].

485 Virus genomes were clustered into viral operational taxonomic units (vOTUs) following
486 MIUViG guidelines (95% average nucleotide identity, ANI; 85% aligned fraction[51]). In brief,
487 we performed an all-versus-all BLAST (version 2.13.0+, '-task megablast -evalue 1e-5 -
488 max_target_seqs 20000') search to estimate pairwise ANIs and AFs, as described in Nayfach et
489 al.[123] (code available at: https://github.com/snayfach/MGV/blob/master/ani_cluster/blastani.py),
490 and employed pyLeiden (available at: https://github.com/apcamargo/pyleiden) to cluster
491 genomes, using as input a graph where pairs of genomes that satisfied the MIUViG criteria were
492 connected by edges. Viruses were also grouped at approximate genus-level (40% AAI —
493 average amino acid identity; 20% shared genes) and family-level (20% AAI; 10% shared genes)
494 clusters using DIAMOND[124] for protein alignment and Markov Cluster Process (MCL)[125] for
495 clustering[52].

496 Viral sequences were assigned to putative host (bacterial and archaeal) taxa through
497 matches to a previously described database of CRISPR spacers of 1.6 million bacterial and
498 archaeal genomes from NCBI GenBank and MAGs (release 242; 15 February 2021)[126–130].
499 Sequences of virus genomes were queried against the spacer database
500 (https://portal.nersc.gov/cfs/m342/crisprDB) using blastn (v2.9.0+, parameters: '-
501 max_target_seqs = 1000 -word_size = 8 -dust = no'). Only alignments with at least 25 bp, less
502 than 2 mismatches, and that covered ≥ 95% of the spacer length were considered. Viral
503 sequences were assigned to the host taxon at the lowest taxonomic rank that had at least two
504 spacers matched and that represented >70% of all matches.

505

506 *Potential Auxiliary Metabolic Gene (AMG) Prediction.*

507 We leveraged an intermediate output of geNomad (version 1.3.3)[122] ('genes.tsv') to
508 screen putative AMGs on the detected viral contigs. Proteins of the viral contigs were annotated
509 by virus and host-specific markers implemented in geNomad. The identified viral hallmark (e.g.,
510 terminase, major capsid protein) and non-hallmark proteins were labeled as 'VV-1' and 'V*-0'
511 in geNomad output, respectively. The rest of the viral proteins of the detected viral contigs that
512 were annotated as non-virus-specific or unclassified were then classified into five categories of
513 putative AMGs based on the presence of viral hallmark or non-hallmarks up- or down-stream as
514 mentioned previously[46]. The AMGs with both virus-specific genes ('VV-1' or 'V*-0') were
515 retained for the following analysis. To improve the functional annotations of the putative AMGs
516 and highlight the viral potentials of metabolizing carbohydrates and glycoconjugates, the AMG
517 proteins were also annotated by Carbohydrate-Active enZYmes (CAZY) Database and Kyoto
518 Encyclopedia of Genes and Genomes (KEGG) database using the default settings in addition to
519 the functional annotation databases implemented in geNomad. The putative AMG was assigned
520 to the functional annotation with the highest bitscore (e.g., duplicate annotations were not
521 allowed). Following Hurwitz and U'Ren[102] and Hurwitz et al.[131], we further screened putative
522 AMGs to remove genes not found in KEGG pathways. Additionally, in recognition of the
523 ambiguity in distinguishing genes encoding auxiliary metabolic functions versus core metabolic
524 processes[102], we discuss the resulting set of genes presented here as 'putative AMGs'

525

526 *Statistical Analysis.*

527

528 All statistical analyses and data visualizations were performed using *R* v 4.1.0[132]. We
529 used the following packages for data manipulation and visualization: ggplot2[133], reshape2[134],
530 pheatmap[135], Hmisc[136], ggpubr[137], RColorBrewer[138], maps[139], stats[140] geosphere[141], plyr[142],
531 dplyr[143], and stringr[144]. Packages pertaining to specific analyses are listed below.
532 We generated rarefaction curves for individual samples (Figure 3) and for cumulative
533 sequencing depth (Figure S1) using the 'phyloseq' package[145] and custom R plots, respectively.
534 Samples containing less than five vOTUs, viral clusters, or viral Pfams; or less than 100
535 CRISPR-spacer-based host taxon, microbial Pfams, or microbial taxa were removed for visual
536 clarity. Removed samples followed the same general trends as presented in Figure 3. To
537 visualize saturation across cumulative sequencing depth (Figure S1), we ordered samples from
538 lowest to highest total assembled bp and progressively added them along the x-axis. On the y-
539 axis, we plot the associated cumulative number of unique attributes.
540 A phylogenetic tree of CRISPR-spacer-based host taxon was generated at the order-level
541 using phyloT v2[146], an online tree generator based on the Genome Taxonomy Database. Then,
542 we visualized the tree in R using the packages 'ggtree'[147], 'treeio'[148], and 'ggnewscale'[149]. To
543 examine relationships between common microbial hosts of soil viruses and soil properties, we
544 first downloaded data describing bulk density, cation exchange capacity, nitrogen, pH, soil
545 organic carbon, and clay content from the SoilGrids250m database[50] using the 'soilDB'
546 package[150]. Mean values of soil properties from 0-5 cm were correlated to the total number of
547 CRISPR spacer hits per microbial ordered using Spearman correlation.
548 Finally, we mapped genes detected across the entire soil virosphere (i.e., all samples
549 combined) to their corresponding KEGG pathways using the 'pathview' package in R[151]. Gene
550 abundances were converted by log base 10 for visualization.

References.

1. Mushegian, A. R. Are there 1031 virus particles on earth, or more, or fewer? *J. Bacteriol.* **202**, (2020).
2. 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).
3. Pratama, A. A. & van Elsas, J. D. The "Neglected" Soil Virome - Potential Role and Impact. *Trends Microbiol.* **26**, 649–662 (2018).
4. 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).
5. Roux, S. & Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond? *Trends Microbiol.* **30**, 1025–1035 (2022).
6. 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).
7. Cobián Güemes, A. G. *et al.* Viruses as winners in the game of life. *Annu. Rev. Virol.* **3**, 197–214 (2016).
8. 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).
9. Emerson, J. B. Soil viruses: A new hope. *mSystems* **4**, (2019).
10. Santos-Medellín, C. *et al.* Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *BioRxiv* (2022) doi:10.1101/2022.03.24.485562.

574 11. Albright, M. B. N. *et al.* Experimental evidence for the impact of soil viruses on carbon
575 cycling during surface plant litter decomposition. *ISME COMMUN.* **2**, 24 (2022).
576 12. Adriaenssens, E. M. *et al.* Environmental drivers of viral community composition in
577 Antarctic soils identified by viromics. *Microbiome* **5**, 83 (2017).
578 13. Durham, D. M. *et al.* Substantial differences in soil viral community composition within
579 and among four Northern California habitats. *ISME COMMUN.* **2**, 100 (2022).
580 14. Liao, H. *et al.* Response of soil viral communities to land use changes. *Nat. Commun.* **13**,
581 6027 (2022).
582 15. Han, L.-L. *et al.* Distribution of soil viruses across China and their potential role in
583 phosphorous metabolism. *Environ. Microbiome* **17**, 6 (2022).
584 16. Braga, L. P. P. *et al.* Impact of phages on soil bacterial communities and nitrogen
585 availability under different assembly scenarios. *Microbiome* **8**, 52 (2020).
586 17. Wu, R. *et al.* Moisture modulates soil reservoirs of active DNA and RNA viruses.
587 *Commun. Biol.* **4**, 992 (2021).
588 18. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat.*
589 *Microbiol.* **3**, 870–880 (2018).
590 19. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.*
591 **5**, 801–812 (2007).
592 20. Keen, E. C. A century of phage research: bacteriophages and the shaping of modern
593 biology. *Bioessays* **37**, 6–9 (2015).
594 21. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat
595 soil. *Microbiome* **9**, 208 (2021).
596 22. Lee, S. *et al.* Methane-derived carbon flows into host-virus networks at different trophic

levels in soil. *Proc Natl Acad Sci USA* **118**, (2021).

23. Starr, E. P. *et al.* Stable-Isotope-Informed, Genome-Resolved Metagenomics Uncovers Potential Cross-Kingdom Interactions in Rhizosphere Soil. *mSphere* **6**, e0008521 (2021).

24. 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).

25. Snyder, J. C. *et al.* Virus movement maintains local virus population diversity. *Proc Natl Acad Sci USA* **104**, 19102–19107 (2007).

26. Hesse, U. *et al.* Virome assembly and annotation: A surprise in the namib desert. *Front. Microbiol.* **8**, 13 (2017).

27. Jin, M. *et al.* Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 58 (2019).

28. Chen, M.-L. *et al.* Viral Community and Virus-Associated Antibiotic Resistance Genes in Soils Amended with Organic Fertilizers. *Environ. Sci. Technol.* **55**, 13881–13890 (2021).

29. Calero-Cáceres, W., Ye, M. & Balcázar, J. L. Bacteriophages as environmental reservoirs of antibiotic resistance. *Trends Microbiol.* **27**, 570–577 (2019).

30. 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).

31. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).

32. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).

33. Williamson, K. E., Radosevich, M., Smith, D. W. & Wommack, K. E. Incidence of

620 lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9**, 2563–
621 2574 (2007).
622 34. Trubl, G. *et al.* Soil viruses are underexplored players in ecosystem carbon processing.
623 *mSystems* **3**, (2018).
624 35. Trubl, G., Hyman, P., Roux, S. & Abedon, S. T. Coming-of-Age Characterization of Soil
625 Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics.
626 *Soil Syst.* **4**, 23 (2020).
627 36. Anantharaman, K. *et al.* Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**,
628 757–760 (2014).
629 37. York, A. Marine microbiology: Algal virus boosts nitrogen uptake in the ocean. *Nat. Rev.*
630 *Microbiol.* **15**, 573 (2017).
631 38. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant
632 ocean viruses. *Nature* **537**, 689–693 (2016).
633 39. Zhang, R., Wei, W. & Cai, L. The fate and biogeochemical cycling of viral elements. *Nat.*
634 *Rev. Microbiol.* **12**, 850–851 (2014).
635 40. Sandaa, R.-A. *et al.* The response of heterotrophic prokaryote and viral communities to
636 labile organic carbon inputs is controlled by the predator food chain structure. *Viruses* **9**,
637 (2017).
638 41. Breitbart, M. Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425–448 (2012).
639 42. Sokol, N. W. *et al.* Life and death in the soil microbiome: how ecological processes
640 influence biogeochemistry. *Nat. Rev. Microbiol.* **20**, 415–430 (2022).
641 43. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev.*
642 *Microbiol.* (2022) doi:10.1038/s41579-022-00811-z.

643 44. Hwang, Y., Rahlff, J., Schulze-Makuch, D., Schloter, M. & Probst, A. J. Diverse viruses
644 carrying genes for microbial extremotolerance in the atacama desert hyperarid soil.
645 *mSystems* **6**, (2021).
646 45. Wu, R. *et al.* Structural characterization of a soil viral auxiliary metabolic gene product - a
647 functional chitosanase. *Nat. Commun.* **13**, 5485 (2022).
648 46. Wu, R. *et al.* DNA Viral Diversity, Abundance, and Functional Potential Vary across
649 Grassland Soils with a Range of Historical Moisture Regimes. *MBio* **12**, e0259521 (2021).
650 47. Liang, X. *et al.* Viral abundance and diversity vary with depth in a southeastern United
651 States agricultural ultisol. *Soil Biol. Biochem* **137**, 107546 (2019).
652 48. Bi, L. *et al.* Diversity and potential biogeochemical impacts of viruses in bulk and
653 rhizosphere soils. *Environ. Microbiol.* **23**, 588–599 (2021).
654 49. Jansson, J. K. Soil viruses: Understudied agents of soil ecology. *Environ. Microbiol.* **25**,
655 143–146 (2023).
656 50. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine
657 learning. *PLoS ONE* **12**, e0169748 (2017).
658 51. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG).
659 *Nat. Biotechnol.* **37**, 29–37 (2019).
660 52. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut
661 microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
662 53. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for
663 interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
664 54. Lindberg, A. A. Bacteriophage receptors. *Annu. Rev. Microbiol.* **27**, 205–241 (1973).
665 55. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z.

666 Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers
667 during the infection process. *Appl. Microbiol. Biotechnol.* **101**, 3103–3119 (2017).
668 56. Kintz, E. *et al.* A BTP1 prophage gene present in invasive non-typhoidal Salmonella
669 determines composition and length of the O-antigen of the lipopolysaccharide. *Mol.*
670 *Microbiol.* **96**, 263–275 (2015).
671 57. Cornell, C. R. *et al.* Temporal Changes of Virus-Like Particle Abundance and
672 Metagenomic Comparison of Viral Communities in Cropland and Prairie Soils. *mSphere* **6**,
673 e0116020 (2021).
674 58. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial
675 and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
676 59. Han, L.-L., Yu, D.-T., Zhang, L.-M., Shen, J.-P. & He, J.-Z. Genetic and functional
677 diversity of ubiquitous DNA viruses in selected Chinese agricultural soils. *Sci. Rep.* **7**,
678 45142 (2017).
679 60. Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome.
680 *Nat. Rev. Microbiol.* **15**, 579–590 (2017).
681 61. Carini, P. *et al.* Relic DNA is abundant in soil and obscures estimates of soil microbial
682 diversity. *Nat. Microbiol.* **2**, 16242 (2016).
683 62. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial
684 diversity. *Nature* **551**, 457–463 (2017).
685 63. Locey, K. J. *et al.* Dormancy dampens the microbial distance-decay relationship. *Philos.*
686 *Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190243 (2020).
687 64. CAZy - GT4. http://www.cazy.org/GT4.html.
688 65. CAZy - GH73. http://www.cazy.org/GH73.html.

689 66. CAZy - CBM50. http://www.cazy.org/CBM50.html.

690 67. Wommack, K. E. & Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems.
691 *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).

692 68. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science* **332**,
693 106–109 (2011).

694 69. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and
695 evolutionary determinants of bacteriophage host range. *Trends Microbiol.* **27**, 51–63
696 (2019).

697 70. Cazares, D. *et al.* A Novel Group of Promiscuous Podophages Infecting Diverse
698 Gammaproteobacteria from River Communities Exhibits Dynamic Intergenus Host
699 Adaptation. *mSystems* **6**, (2021).

700 71. De Sordi, L., Khanna, V. & Debarbieux, L. The gut microbiota facilitates drifts in the
701 genetic diversity and infectivity of bacterial viruses. *Cell Host Microbe* **22**, 801-808.e3
702 (2017).

703 72. Cornuault, J. K. *et al.* The enemy from within: a prophage of Roseburia intestinalis
704 systematically turns lytic in the mouse gut, driving bacterial adaptation by CRISPR spacer
705 acquisition. *ISME J.* **14**, 771–787 (2020).

706 73. Heineman, R. H., Springman, R. & Bull, J. J. Optimal foraging by bacteriophages through
707 host avoidance. *Am. Nat.* **171**, E149-57 (2008).

708 74. Meyer, J. R. *et al.* Repeatability and contingency in the evolution of a key innovation in
709 phage lambda. *Science* **335**, 428–432 (2012).

710 75. Ford, B. E. *et al.* Frequency and fitness consequences of bacteriophage φ6 host range
711 mutations. *PLoS ONE* **9**, e113078 (2014).

712 76. Holtzman, T. *et al.* A continuous evolution system for contracting the host range of
713 bacteriophage T7. *Sci. Rep.* **10**, 307 (2020).
714 77. Sant, D. G., Woods, L. C., Barr, J. J. & McDonald, M. J. Host diversity slows
715 bacteriophage adaptation by selecting generalists over specialists. *Nat. Ecol. Evol.* **5**, 350–
716 359 (2021).
717 78. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
718 79. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches
719 to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
720 80. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and its
721 constraints on virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
722 81. Srinivasiah, S. *et al.* Phages across the biosphere: contrasts of viruses in soil and aquatic
723 environments. *Res. Microbiol.* **159**, 349–357 (2008).
724 82. Srinivasiah, S. *et al.* Dynamics of autochthonous soil viral communities parallels dynamics
725 of host communities under nutrient stimulation. *FEMS Microbiol. Ecol.* **91**, (2015).
726 83. Van Goethem, M. W., Swenson, T. L., Trubl, G., Roux, S. & Northen, T. R.
727 Characteristics of Wetting-Induced Bacteriophage Blooms in Biological Soil Crust. *MBio*
728 **10**, (2019).
729 84. Lourenço, M. *et al.* The spatial heterogeneity of the gut limits predation and fosters
730 coexistence of bacteria and bacteriophages. *Cell Host Microbe* **28**, 390-401.e5 (2020).
731 85. Mattenberger, Y., Mattson, S., Métrailler, J., Silva, F. & Belin, D. 55.1, a gene of unknown
732 function of phage T4, impacts on Escherichia coli folate metabolism and blocks DNA
733 repair by the NER. *Mol. Microbiol.* **82**, 1406–1421 (2011).
734 86. Romine, M. F. *et al.* Elucidation of roles for vitamin B12 in regulation of folate,

735 ubiquinone, and methionine metabolism. *Proc Natl Acad Sci USA* **114**, E1205–E1214
736 (2017).
737 87. Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L. & Davidson, A. R. The phage
738 lambda major tail protein structure reveals a common evolution for long-tailed phages and
739 the type VI bacterial secretion system. *Proc Natl Acad Sci USA* **106**, 4160–4165 (2009).
740 88. Records, A. R. The type VI secretion system: a multipurpose delivery system with a
741 phage-like machinery. *Mol. Plant Microbe Interact.* **24**, 751–757 (2011).
742 89. Kanamaru, S. Structural similarity of tailed phages and pathogenic bacterial secretion
743 systems. *Proc Natl Acad Sci USA* **106**, 4067–4068 (2009).
744 90. Sgro, G. G. *et al.* Bacteria-Killing Type IV Secretion Systems. *Front. Microbiol.* **10**, 1078
745 (2019).
746 91. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a
747 DNA substrate. *Science* **304**, 1170–1173 (2004).
748 92. Lasker, K., Mann, T. H. & Shapiro, L. An intracellular compass spatially coordinates cell
749 cycle modules in Caulobacter crescentus. *Curr. Opin. Microbiol.* **33**, 131–139 (2016).
750 93. Tan, Y. *et al.* Organic fertilizers shape soil microbial communities and increase soil amino
751 acid metabolites content in a blueberry orchard. *Microb. Ecol.* **85**, 232–246 (2023).
752 94. Wang, D., Chadwick, D. R., Hill, P. W., Ge, T. & Jones, D. L. Rapid microbial uptake and
753 mineralization of 14C-labelled cysteine and methionine along a grassland productivity
754 gradient. *Soil Biol. Biochem* **180**, 109022 (2023).
755 95. Ma, Q. *et al.* Carbon and sulphur tracing from soil organic sulphur in plants and soil
756 microorganisms. *Soil Biol. Biochem* **150**, 107971 (2020).
757 96. Wang, D., Wang, J., Chadwick, D. R., Ge, T. & Jones, D. L. Competition for two sulphur

758 containing amino acids (cysteine and methionine) by soil microbes and maize roots in the
759 rhizosphere. *Biol. Fertil. Soils* **59**, 697–704 (2023).
760 97. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the
761 dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**,
762 892–899 (2007).
763 98. Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. *Annu. Rev. Genet.* **42**,
764 647–681 (2008).
765 99. Jia, K. *et al.* A novel inovirus reprograms metabolism and motility of marine alteromonas.
766 *Microbiol. Spectr.* **10**, e0338822 (2022).
767 100. Wu, D., Yuan, Y., Liu, P., Wu, Y. & Gao, M. Cellular responses in Bacillus thuringiensis
768 CS33 during bacteriophage BtCS33 infection. *J. Proteomics* **101**, 192–204 (2014).
769 101. Kühlbrandt, W. Structure and Mechanisms of F-Type ATP Synthases. *Annu. Rev.*
770 *Biochem.* **88**, 515–549 (2019).
771 102. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems.
772 *Curr. Opin. Microbiol.* **31**, 161–168 (2016).
773 103. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine
774 microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
775 104. Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. Shedding new light on viral
776 photosynthesis. *Photosyn. Res.* **126**, 71–97 (2015).
777 105. Daniel, R. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).
778 106. Leite, M. F. A., van den Broek, S. W. E. B. & Kuramae, E. E. Current challenges and
779 pitfalls in soil metagenomics. *Microorganisms* **10**, (2022).
780 107. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T.

781 Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* **3**,
782 (2018).
783 108. Epstein, R. H., Bolle, A. & Steinberg, C. M. Amber mutants of bacteriophage T4D: their
784 isolation and genetic characterization. *Genetics* **190**, 833–840 (2012).
785 109. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil.
786 *Methods Mol. Biol.* **501**, 15–21 (2009).
787 110. Wommack, K. E., Williamson, K. E., Helton, R. R., Bench, S. R. & Winget, D. M.
788 Methods for the isolation of viruses from environmental samples. *Methods Mol. Biol.* **501**,
789 3–14 (2009).
790 111. Trubl, G. *et al.* Towards optimized viral metagenomes for double-stranded and single-
791 stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
792 112. Zablocki, O. *et al.* VirION2: a short- and long-read sequencing and informatics workflow
793 to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088 (2021).
794 113. Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and
795 microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800
796 (2019).
797 114. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and
798 abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
799 115. Guerra, C. A. *et al.* Blind spots in global soil biodiversity and ecosystem function research.
800 *Nat. Commun.* **11**, 3870 (2020).
801 116. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.8: overview and updates.
802 *Nucleic Acids Res.* **49**, D723–D733 (2021).
803 117. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative

804 analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–
805 D677 (2019).
806 118. Clum, A. *et al.* DOE JGI metagenome workflow. *mSystems* **6**, (2021).
807 119. Nelson, W. C. *et al.* Terabase metagenome sequencing of grassland soil microbiomes.
808 *Microbiol. Resour. Announc.* **9**, (2020).
809 120. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**,
810 D427–D432 (2019).
811 121. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer
812 based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*
813 **5**, 69 (2017).
814 122. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nature*
815 *Biotechnology* (2023) doi:10.1038/s41587-023-01953-y.
816 123. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-
817 assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
818 124. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
819 using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
820 125. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM. J. Matrix Anal.*
821 *& Appl.* **30**, 121–141 (2008).
822 126. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–
823 509 (2021).
824 127. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
825 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
826 128. 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).

129. Merrill, B. D. *et al.* Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers Vanishing Gut Microbes. *BioRxiv* (2022) doi:10.1101/2022.03.30.486478.

130. 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).

131. 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).

132. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2021).

133. Wickham, H. & Chang, W. Package "ggplot2." *Create elegant data …* (2016).

134. Wickham, H. reshape2: flexibly reshape data: a reboot of the reshape package.

135. Kolde, R. & Kolde, M. R. Package "pheatmap." *R Package* (2018).

136. Jr, F. H. & Jr, M. H. Package "hmisc." *CRAN2018* (2019).

137. Goslee, S. C. & Urban, D. L. The ecodist Package for Dissimilarity-based Analysis of Ecological Data. *J. Stat. Softw.* **22**, (2007).

138. Neuwirth, E. & Brewer, R. C. ColorBrewer palettes. *R package version* (2014).

139. Becker, R. A., Wilks, A. R., Brownrigg, R. & Minka, T. P. maps: Draw geographical maps.

140. R: The R Stats Package. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.

850 141. Hijmans, R. J., Williams, E., Vennes, C. & Hijmans, M. R. J. Package "geosphere."
851 142. Wickham, H. Tools for Splitting, Applying and Combining Data [R package plyr version
852 1.8.8]. (2022).
853 143. A Grammar of Data Manipulation [R package dplyr version 1.1.2]. https://cran.r-
854 project.org/web/packages/dplyr/index.html (2023).
855 144. Wickham, H. Simple, Consistent Wrappers for Common String Operations [R package
856 stringr version 1.5.0]. (2022).
857 145. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive
858 analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
859 146. phyloT : a phylogenetic tree generator. https://phylot.biobyte.de/.
860 147. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: An R package for
861 visualization and annotation of phylogenetic trees with their covariates and other
862 associated data. *Methods Ecol. Evol.* **8**, 28–36 (2016).
863 148. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with
864 Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
865 149. Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. *R package version*
866 *0.4* (2020).
867 150. DE Beaudette, Skovlin, J., Roecker, S. & Beaudette, M. D. E. Package "soilDB." (2022).
868 151. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data
869 integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).
870 152. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in
871 nature: mechanisms, impact and ecology of temperate phages. *ISME J.* **11**, 1511–1520
872 (2017).

153. Koskella, B. & Brockhurst, M. A. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* **38**, 916–931 (2014).

154. Knowles, B. W. *Piggyback-the-Winner: lytic to temperate switching of viral communities*. (search.proquest.com, 2016).

155. Maslov, S. & Sneppen, K. Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. *Sci. Rep.* **7**, 39642 (2017).

156. Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–351 (2012).

157. LTER SOM Database. https://cosima.nceas.ucsb.edu/lter-som/.

158. Wieder, W. R. *et al.* SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0. *Earth Syst. Sci. Data* **13**, 1843–1854 (2021).

159. Suttle, C. A. The significance of viruses to mortality in aquatic microbial communities. *Microb. Ecol.* **28**, 237–243 (1994).

160. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).

161. Rodriguez-Brito, B. *et al.* Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**, 739–751 (2010).

162. Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A. & Giske, J. A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc Natl Acad Sci USA* **111**, 7813–7818 (2014).

896 163. Liang, X. & Radosevich, M. Commentary: A Host-Produced Quorum-Sensing
897 Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Front. Microbiol.* **10**, 1201
898 (2019).
899 164. Weinbauer, M. G. & Rassoulzadegan, F. Are viruses driving microbial diversification and
900 diversity? *Environ. Microbiol.* **6**, 1–11 (2004).
901 165. De Smet, J. *et al.* High coverage metabolomics analysis reveals phage-specific alterations
902 to Pseudomonas aeruginosa physiology during infection. *ISME J.* **10**, 1823–1835 (2016).
903 166. Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through
904 diverse mechanisms. *ISME J.* **10**, 2854–2866 (2016).
905 167. Dey, P. *et al.* Physiological, biochemical, and molecular adaptation mechanisms of
906 photosynthesis and respiration under challenging environments. in *Plant perspectives to*
907 *global climate changes* 79–100 (Elsevier, 2022). doi:10.1016/B978-0-323-85665-2.00003-
908 0.
909 168. Huchzermeyer, B., Hausmann, N. & Paquet-Durant, F. Biochemical and physiological
910 mechanisms leading to salt tolerance. *Tropical …* (2004).