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 soil metagenomes that is comprised of 616,935 uncultivated viral 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 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 soil compared to other ecosystems[1,8–10], as well as the high diversity of their microbial hosts. Indeed, soil viral abundance and composition vary with factors such as 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 microbial communities[8,10,19]. For instance, spatial viral turnover has been shown to be over 5 times higher than microbial community turnover across an 18m 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, physicochemical interactions, host organisms, and food web implications of the soil virosphere. 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 metagenomes to date, encompassing contributions from prominent repositories, ecological networks, and individual collaborators.

RESULTS

76 *Global Soil Virus (GSV) Atlas.*

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

88 We also collected associated environmental parameters describing each sample from the
89 SoilGrids250m database[33]. We assayed a wide variety of soils that ranged from bulk density of
90 0.24-1.56 (kg/dm^3), cation exchange capacity of 6.8-71 (CEC; cmol(c)/kg), nitrogen content of
91 0.19-22.4 (N; g/kg), pH of 4.3-8.5, soil organic C of 1.9-510.9 (SOC; g/kg), and clay content of
92 2.7-57.1 (%) (Fig. 1).

93 The 49,649 UViGs of sufficient quality for downstream analysis clustered into 38,508
94 vOTUs at the species-like level[34], of which 3,296 were previously unrepresented in IMG/VR
95 (Fig. 2a-b). Only 13.9% of the GSV Atlas' vOTUs appeared in more than one sample, and less
96 than 1% were present in more than 5 samples. At higher taxonomic levels, we found 21,160 and
97 7,598 clusters at the genus and family level respectively[35]. This equates to an average of 40.01
98 (range: 1 - 2,124), 35.48 (range: 1 - 1,651), and 24.91 (range: 1 - 896) unique viral clusters per

99 sample at the species, genus, and family levels. 38,278 out of 38,508 vOTUs (99.4%) had at least
100 one member assigned to a taxon by geNomad.

101 We identified 1,432,147 viral genes, of which only 260,258 (~18%) were found in at
102 least one annotation database (1,022, 3,634, and 145 unique KO, Pfam, and CAZy annotations,
103 Fig. 2). After filtering to putative AMGs (see methods)[30], we found 5,043 genes that mapped to
104 83 KEGG pathways (1,941 KO and 3,357 CAZyme, some genes were associated with both KO
105 and CAZyme). The median per sample putative AMG abundance (gene copies per sample) was
106 ~19 genes (median = 4, reflecting skewing from a few samples with high AMG abundance). The
107 most abundant KEGG pathways containing putative AMGs were Galactose metabolism,
108 Glycosaminoglycan degradation, Starch and sucrose metabolism, N-Glycan biosynthesis, and
109 Folate biosynthesis; while the most abundant gene annotations associated with AMGs were
110 Glycosyl Transferase Family 4 (GT4), Carbohydrate-Binding Module Family 50 (CMB50), 6-
111 pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase (K01737), Glycoside Hydrolase
112 Family 73 (GH73), and 7-carboxy-7-deazaguanine synthase (K10026) (Fig. 2c-d).

113 In contrast to the saturation observed in rarefaction curves for microbial taxonomy and
114 genes annotated by Pfam, rarefaction curves of soil viruses for individual samples (vOTUs and
115 and viral clusters) and their genes (annotated by Pfams) did not reach an asymptote (Fig. S1).
116 The number of new and unique vOTUs and viral clusters at the family level (Fig. S1b and S1c)
117 was linearly related to sequencing depth, while viral Pfams displayed slight curvature. These
118 linear relationships were observed when considering 2TB of metagenomic sequencing—4-fold
119 more sequencing depth than any other soil metagenome in this study and 40-fold more than the
120 Joint Genome Institute (JGI) recommended sequencing depth for soil samples (45 GB). When

121 considering cumulative unique attributes versus sequencing depth (Fig. S2), relationships in
122 vOTUs and viral clusters displayed slight curvature, while viral Pfams neared saturation.
123
124 *Microbial Hosts of Soil Viruses.*
125 We connected 1,450 viruses to putative hosts of 82 different bacterial and archaeal orders
126 with CRISPR spacers. This equates to 2.78% of QA/QC-ed viral contigs that were associated
127 with CRISPR spacer hits; roughly 70% more host assignments than in another recent
128 assessment[4,36]. 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 1,223 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 microbial order.
139 The most prevalent host taxa were distributed across distantly-related phyla, including
140 members of prominent soil orders such as Pseudomonadales, Burkholderiales, Acidobacteriales,
141 and Bacteroidales (Fig. 3b). The frequency of CRISPR hits associated with Acidobacteriales,
142 Oscillospirales, Pedosphaerales, and Geobacterales were positively correlated to soil nitrogen,
143 organic carbon, and CEC; while Enterobacterales, Obscuribacterales, Mycobacteriales,

144 Pseudomonadales, and Streptomycetales were positively correlated to soil bulk density, and to a
145 lesser extent, pH and clay.
146
147 *Metabolic potential encoded by the soil virosphere.*
148 Because viral gene annotations were sparsely distributed across many functions, we
149 screened all viral genes (regardless of assignment by the AMG pipeline) against KEGG
150 pathways to understand relationships among genes in the context of known metabolic processes.
151 Across the entire soil virosphere, we uncovered portions of KEGG pathways that were mostly
152 complete, including functions involved in amino acid and sugar metabolism and biosynthesis,
153 antibacterial mechanisms, nucleotide synthesis, and other viral functions (e.g., infection
154 strategies, Supplemental Material and Fig. 4). Of note, we found nearly complete portions of
155 energy-generating pathways including the pentose phosphate pathways and F-type ATPase-
156 mediated portions of photosynthesis. Lipopolysaccharide pathway-related genes (LPS) that may
157 be important as host receptors for bacteriophage and prevention of superinfection were also
158 prevelant[37,38].
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. Viral diversity in the GSV Atlas appeared to be largely distinct
164 from other global habitats. Nearly 80% of GVS Atlas sequences that clustered with existing
165 sequences in IMG/VR were attributed to soil or soil-like habitats [i.e. 'other terrestrial' or 'plant-
166 associated' (rhizosphere)], underscoring the unique composition of the soil virosphere relative to

167 more well-studied marine and human environments. Additionally, few shared vOTUs and viral
168 clusters between samples may indicate high spatial turnover (i.e., changes in soil virosphere
169 composition through space). Recent studies have estimated soil viral diversity is high both
170 relative to other viral habitats and relative to soil microbial diversity[7,8,10,22]. However, these
171 estimates have been limited by copious viral and microbial 'dark matter' for which no functional
172 or taxonomic assignment is known[14,23,32]. Towards this end, the diversity encompassed by the
173 GSV Atlas can serve as a community resource for characterizing this unknown fraction of the
174 soil virosphere.

175 Analysis of the GSV Atlas suggests that extreme spatial heterogeneity may be a key
176 feature of the soil virosphere at the global scale. While rapid viral spatial turnover was recently
177 observed across short spatial scales (<10-20m)[8,10], there has been no such demonstration of viral
178 biogeography across the world. We propose that high rates of spatial turnover could result from
179 low dispersal rates or distinct temporal dynamics of viral communities relative to other
180 organisms. For example, while dormant microorganisms and relic DNA can persist for months or
181 more[39–41], the burst of viral cells associated with active infections may generate short-lived
182 pulses of distinct viral communities that do not contribute to relic DNA due to their
183 comparatively small genome sizes versus microorganisms. Additionally, the apparent
184 discrepancies between microbial and viral dispersal processes could be due to the presence of
185 free viruses that are not actively involved in microbial infection[14], smaller viral genomes that
186 could facilitate physical protection, differences in traits that facilitate dispersal between viruses
187 and microorganisms, variation in bioinformatic pipelines, and/or other ecological differences
188 between viruses and microorganisms.

189 Together, these factors make characterizing the soil virosphere an immense challenge for
190 the coming decade. When examining individual soil samples, the number of new and unique
191 viral attributes (e.g., vOTUs, clusters, Pfams) was linearly related to sequencing depth,
192 suggesting that new viral discoveries are likely to continue with increasingly deep sequencing
193 (Fig. S1). This contrasts with rarefaction curves of the soil microbiome and of microbial hosts of
194 soil viruses, which both asymptoted well before sequencing depths of typical soil microbial
195 investigations. Still, when looking at the cumulative number of unique viral attributes detected in
196 all samples collectively (Fig. S2), many viral attributes began to saturate with sequencing depth.
197 This suggests that while individual samples do not capture soil viral diversity, we can begin to
198 constrain the extent of diversity when sequences from thousands of existing samples are
199 aggregated.
200 Assembling the GSV Atlas enables us to begin to describe the enormous functional
201 diversity in the soil virosphere. After filtering viral genes to putative AMGs (see methods), some
202 KEGG pathways represented by the most putative AMGs were associated with major soil carbon
203 cycling processes (map00052, Galactose metabolism; map00500, Starch and sucrose
204 metabolism). Likewise, at the level of gene annotations, the most common putative AMGs
205 suggested a role for viruses in soil carbon cycles. Glycosyltransferases 4 (GT4) and
206 glycosylhydrolases 73 (GH73) encode enzymes involved in the metabolism and/or production of
207 sugars common to soils including sucrose, mannose, glucosamine, and maltose[42,43]; and
208 Carbohydrate Binding Module 50 (CBM50) has been previously linked to chitin – one of the
209 most abundant carbon molecules in soil[44].
210 Collectively, these genes encode the decomposition of chemically complex organic
211 material for microbial energy generation and growth. Our results are consistent with previous

212 work from single locations that have hinted at a wide range of possible soil viral AMGs,
213 including glycoside hydrolases, carbohydrate esterases, and carbohydrate-binding modules[15,23,31].
214 Given that a large proportion of soil microorganisms are infected by viruses at any given time[45],
215 AMGs encoded by soil viruses have the potential to impact global biogeochemical
216 cycles[15,22,23,31]. The thousands of putative AMGs identified here represent the most extensive
217 survey to date and further impress the importance of the soil virosphere as a reservoir for
218 biogeochemical potential.
219
220 *Microbial Hosts of Soil Viruses.*
221 Unraveling relationships between viruses and their host communities is imperative to
222 understanding the impact of the virosphere on soil processes. Host presence should be tightly
223 coupled to viral abundance, and in turn, these linkages are mediated by spatial, temporal, and
224 environmental factors[15,46,47]. These linkages are also dependent on viral host range (i.e., host
225 specificity) — higher host specificity should lead to stronger coupling between microbial and
226 viral abundance and community composition. Viral host specificity is also associated with
227 ecological factors that impact microbial community composition and may result in tradeoffs
228 between viral growth and the breadth of the host range[11,48–50]. These dynamics are unresolved in
229 soils, leading to open questions regarding associations between viral host specificity and
230 microbial community functions.
231 Our analysis of the GSV Atlas indicates that limited host infection range may be a
232 dominant trait across soil viruses. Across the GSV Atlas, there were few hosts per vOTU on
233 average (mean = 0.42), and of vOTUs associated with multiple host sequences, the vast majority
234 were linked to multiple hosts of the same phylogenetic clade. While high host specificity has

235 historically been the prevailing paradigm, our work contrasts recent studies suggesting that some
236 soil viruses may have broader host ranges than viruses in other habitats[51,52].
237 As a whole, we identified taxonomically diverse soil microorganisms as viral hosts (Fig.
238 3). Host sequences spanned nearly every major soil microbial clade, supporting recent studies
239 that detected soil viral hosts classified as Acidobacteria, Verrucomicrobia, Deltaproteobacteria,
240 and Gammaproteobacteria[22,23,31]. The ultimate impact of viral predation on soil functions is at
241 least partially associated with the taxonomic distribution of hosts—viral impacts may be larger
242 when the hosts contribute to key ecosystem functions. The taxonomic breadth of the microbial
243 hosts of soil viruses also suggests a role for the soil virosphere in most soil habitats on Earth, as
244 the microbial hosts found here occupy a wide range of ecological niches. Moreover, some hosts
245 were susceptible to changes in the environment, underscoring the role that soil heterogeneity
246 may play both in viral diversity and in shaping viral-host dynamics.
247 Relationships between environmental factors and host taxa may reflect environmental
248 filtering on host communities (which in turn, determine the amount and type of viruses present)
249 or on viruses directly which subsequently impacts host community composition[15,27,53,54]. Viral
250 infections have been previously linked to soil parameters including moisture[12,30] and carbon and
251 nitrogen content[9]. In our analysis, bulk density may serve as a proxy for hydrologic connectivity
252 in the soil matrix. For example, low hydrologic connectivity may create 'spatial refuges' for soil
253 bacteria from viral infections[8], influence the virus-host encounter rates, and thus structure the
254 soil virosphere and its hosts. Nutrient amendments are also considered to be drivers of the soil
255 virosphere, supporting the relationship we observed between carbon, nitrogen, and host taxa.
256
257 *Metabolic Potential Encoded by the Soil Virosphere.*

258 We detected many hallmarks of viral activity in the soil virosphere, including genes
259 associated with cell lysis, DNA repair/replication, and other infection signatures. For instance,
260 across the entire soil virosphere, genes associated with DNA mismatch repair (map03430),
261 homologous recombination (map03440), and base excision repair (map03410) were prevalent
262 (Fig. S3). Folate biosynthesis was also common in the soil virosphere. Folate and other B
263 vitamins are key to bacterial growth (map00670, map00790, Fig. 4)[55] and therefore may be
264 logical targets for pathogens[56]. Bacterial secretion systems (map03070, Fig. S4), which may be
265 evolutionarily derived from phages[57], were also found in many soil viruses. Type IV secretion
266 systems in particular can be used by bacteria to secrete toxins[58] or as a method for DNA transfer
267 through membranes[59]. Finally, the *Caulobacter* cell cycle (map04112, Fig. S5), which has a
268 distinct division pattern[60], was rife across soils and may be a promising indicator of viral
269 infections. The prevalence of viral genes associated with central microbial functions highlights
270 the potential importance of viral activity in soils and the need for targeted approaches to quantify
271 the extent and impact of viral gene expression.

272 The GSV Altas also contains many viral amino acid biosynthesis/degradation pathways
273 that could be critical in viral life cycles(e.g., map00250, map00260, map00270, map00330,
274 map00340, Fig. S6). Amino acids are building blocks for cellular material and also support soil
275 biogeochemical cycles, as they can enhance carbon cycling through priming effects and/or
276 enhanced nutrient availability[61,62] (e.g., map00500, map00052, map00051, Fig. 4). Collectively,
277 these pathways demonstrate several possibilities for soil viral impacts on processes that are
278 central to microbial metabolism and biogeochemical 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 (map00190), pentose phosphate pathway

281 (map00030), and LPS (map00540). Five of seven subcomponents of the F-type ATPase were
282 detected in the soil virosphere, while no V- or A-type ATPases were found. Given the
283 evolutionary similarities between V- and F-type ATPase in particular[63], the lack of any V- or A-
284 type ATPase components is notable in light of the near complete F-type ATPase. Though there is
285 some basis for F-type ATPases in viral replication[64], we also note the possible involvement of F-
286 type ATPase in photosynthetic energy generation[65]. Given the prominence of photosynthetic
287 marine AMGs[26,66], we highlight the possibility of a viral F-type ATPase as a soil AMG. The
288 pentose phosphate pathway is also a prevalent and important AMG found in marine ecosystems,
289 where viral infection diverts carbon towards the pentose phosphate pathway as an 'express route'
290 of energy generation, at the expense of host carbon metabolism (reviewed in [66]). Finally, we
291 observed nearly complete LPS pathways in the GSV Atlas. Phages often carry depolymerases
292 and other enzymes that target LPS or similar outer membrane components to facilitate binding
293 and entry[37]. However, the representation of the LPS synthesis pathway by putative soil AMGs
294 indicates that phage may work to change the function of the pathway post-infection, potentially
295 to prevent superinfection[38]. Collectively, we propose that F-type ATPase, pentose phosphate
296 pathway, and LPS may be interesting pathways for more targeted investigations into the role of
297 the virosphere in soil microbiome function.

298

299 *Future Recommendations.*

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

304 ecologists and suggestions for surmounting them.

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

317 Knowledge about soil viral diversity and function is also limited by gaps in field and
318 laboratory experiments. The GSV Atlas demonstrates that extensive, spatially-explicit sampling
319 is needed to capture the high spatial turnover of the soil virosphere. The spatial coverage of most
320 'global' ecological studies, including this one, often suffers from large data gaps[73]. Concerted
321 efforts are needed to sample wide spatial domains, including historically undersampled regions,
322 given the high viral diversity uncovered by the GSV Atlas. Expansion of the known virosphere
323 in this way will also help to facilitate tool development. Although we did not assess temporal
324 dynamics, temporally-explicit approaches are likewise needed to characterize temporal dynamics
325 in soil viral communities. Further, our functional annotation of viral contigs revealed diverse
326 genes associated with functions relevant to both viral and microbial communities, and it is

327 impossible to know the true functions of viral genes without targeted functional assays. We
328 therefore propose that experiments targeting the expression and auxiliary metabolic function of
329 viral genes are needed to properly assess AMGs in viral communities.

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

338

339 *Conclusions.*

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

Correspondence should be addressed to emily.graham@pnnl.gov.

Acknowledgements. This work was supported by the U.S. Department of Energy (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 U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We thank Young Song for his help in data visualization.

Author Contributions. EBG, RW, JKJ, DPE, NCK, and JEM conceived of this project. EBG, RW, APC, RYN, MN, and AC conducted data analysis. EBG, RW, APC, RYB, NCK, JKJ, KSH, and JEM 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

367 **FIGURES**

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

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

379 **Fig. 3. Relationships between soil viruses and their hosts. a,** Cumulative Distribution
380 Function plot showing the ratio of CRISPR spacer hits to viral sequences per sample. Inset
381 shows a zoomed portion of the plot from 0 to 0.08 along the y-axis. **b,** Phylogenetic tree of
382 bacterial hosts at the order level. Phylum taxonomy is shown in the inner circle and the
383 abundance of CRISPR spacer hits to each order is shown in the outer circle. Two archaeal orders
384 (Nitrososphaerales and Halobacteriales) of hosts are not shown. **c**, Correlation between
385 frequency of CRISPR hits (defined as total CRISPR spacer hits per microbial order) and
386 environmental parameters from SoilGrids250m. Color denotes Spearman rho. For inclusion in
387 the heatmap, host order must be present in more than 5 samples.

388 **Fig. 4. Metabolic potential encoded by the soil virosphere.** A cellular diagram depicting
389 portions of the F-type ATPase (map00190), lipopolysaccharide pathway (map00540), pentose

390 phosphate pathway (map00030), and vitamin B- and amino acid-related KEGG pathways in the
391 soil virosphere. Genes detected in the soil virosphere are colored according to pathway type.
392 Undetected genes and associated metabolites (unmeasured) are grayed out. Arrow width deontes
393 gene abundance.
394

395 **METHODS**

396 *Data Collection and Curation.*

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

403 For samples collected via JGI IMG/M, we retrieved all studies with GOLD[74] ecosystem
404 type of "Soil" as of August 2020. We manually curated metagenomic sequences to remove
405 misclassified data as follows. We removed samples from studies with the following: (1) GOLD
406 ecosystem types: Rock-dwelling, Deep subsurface, Plant litter, Geologic, Oil reservoir, Volcanic,
407 and Contaminated; (2) GOLD ecosystem subtypes: Wetlands, Aquifer, Tar, Sediment, Fracking
408 Water, and Soil crust; (4) words in title: wetland, sediment, acid mine, cave wall surface, mine
409 tailings, rock biofilm, beach sand, Petroleum, Stalagmite, Subsurface hydrocarbon microbial
410 communities, Vadose zone, mud volcano, Fumarolic, enriched, Composted filter cake, Ice
411 psychrophilic, oil sands, groundwater, Contaminated, rock biofilm, Deep mine, coal mine fire,
412 Hydrocarbon resource environments, Marine, enrichment, groundwater, mangrove, saline desert,
413 Hydroxyproline, Rifle, coastal, compost, biocrust, crust, Creosote, soil warming, Testing DNA
414 extraction, and/or Agave; (5) GOLD geographic location of wetland; and (6) GOLD project type
415 of Metagenome - Cell Enrichment. Additionally, sample names that indicated experimental
416 manipulation (e.g., CO_2 enrichment or nitrogen fertilization) or were located in permafrost layers

417 were manually excluded. This resulted in 1,480 curated metagenomes from publicly available
418 data in IMG/M.
419 After collating samples from JGI IMG/M and the newly collected samples from external
420 networks and collaborators, the final dataset consisted of 2,953 soils with 2,015,688,128 contigs,
421 representing 1.2 terabases of assembled DNA sequences.
422 In parallel, we retrieved mean values for soil parameters from the SoilGrids250m
423 database from 0-5 cm[33]. SoilGrids250m is a spatial interpolation of global soil parameters using
424 ~150,000 soil samples and 158 remote sensing-based products. Here, we focus on six
425 parameters often associated with soil microbial communities: bulk density, cation exchange
426 capacity, nitrogen, pH, soil organic carbon, and clay content. Because our focus on spatial
427 dynamics and soils were collected at various times, we did not include temporally dynamic
428 variables such as soil moisture or temperature in our set of environmental parameters, though we
429 acknowledge they may have profound impacts on the soil virosphere.
430
431 *Assembly and Annotation of Samples added to IMG/M.*
432 To standardize data analysis across all samples, the 1,552 soil metagenomic samples not
433 collected from IMG/M were analyzed using the JGI's Metagenome Workflow [75]. In brief,
434 samples were individually assembled using MetaSpades v. 3.1. 1,476 of the 1,552 assembled soil
435 samples passed default quality control thresholds[76], yielding 133 gigabases of assembled DNA in
436 241,465,924 contigs. Additionally, three very large metagenomes (>1TB each) were assembled
437 separately due to computational limitations in standard workflows[77]. The resulting assemblies
438 were assigned GOLD identification numbers and imported into IMG/M and processed using

439 version 5.0.0 of the IMG/M Metagenome Annotation Pipeline to align with data obtained
440 directly from IMG/M[75].
441
442 *Virus identification, clustering, and host prediction.*
443 We performed an initial identification of viral contigs using using a modified version of
444 the IMG/VR v3's virus identification pipeline (code available at
445 https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline)[35,36]. The pipeline
446 identifies viruses based on the presence of 23,841 virus protein families, 16,260 protein families
447 of microbial origin from the Pfam database[78] and VirFinder[79] to identify putative viral genomes
448 in contigs that were at least 1 kb long. During the course of this study, geNomad (version
449 1.3.3)[80], a tool for virus identification with improved classification performance was released and
450 incorporated into our pipeline to improve prediction confidence and perform taxonomic
451 assignment. We further processed predicted viral sequences using CheckV v1.0.1 (database
452 version 1.5)[81] to assess the quality of the viral genomes. As this study focused on non-integrated
453 virus genomes, contigs that were flagged by either geNomad or CheckV as proviral were
454 discarded. From the remaining contigs, virus genomes were selected using the following rules:
455 (1) contigs of at least 1kb with high similarity to genomes in the CheckV database (that is, that
456 had high- or medium-quality completeness estimates) or that contained direct terminal repeats
457 were automatically selected; (2) contigs longer than 10kb were required to have a geNomad virus
458 score higher than 0.8 and to either encode one virus hallmark (e.g., terminase, capsid proteins,
459 portal protein, etc.), as determined by geNomad, or to have a geNomad virus marker of at least
460 5.0; (3) contigs shorter than 10 kb and longer than 5 kb were required to have a geNomad virus
461 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: https://doi.org/10.25584/2229733.

Viral genomes were clustered into viral operational taxonomic units (vOTUs) following MIUViG guidelines (95% average nucleotide identity, ANI; 85% aligned fraction[34]). In brief, we performed an all-versus-all BLAST (version 2.13.0+, '-task megablast -evalue 1e-5 -max_target_seqs 20000') search to estimate pairwise ANIs and AFs, as described in Nayfach et al.[81] and employed pyLeiden (available at: https://github.com/apcamargo/pyleiden) 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% AAI — average amino acid identity; 20% shared genes) and family-level (20% AAI; 10% shared genes) clusters using DIAMOND[82] for protein alignment and Markov Cluster Process (MCL)[83] 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)[84–88]. Sequences of viral genomes were queried against the spacer database (https://portal.nersc.gov/cfs/m342/crisprDB) using blastn (v2.9.0+, parameters: '-max_target_seqs = 1000 -word_size = 8 -dust = no'). Only alignments with at least 25 bp, less than 2 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.

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

486 We leveraged an intermediate output of geNomad (version 1.3.3)[80] ('genes.tsv') to screen
487 putative AMGs on the detected viral contigs. Proteins of the viral contigs were annotated by
488 virus and host-specific markers implemented in geNomad. The identified viral hallmark (e.g.,
489 terminase, major capsid protein) and non-hallmark proteins were labeled as 'VV-1' and 'V*-0'
490 in geNomad output, respectively. The rest of the viral proteins of the detected viral contigs that
491 were annotated as non-virus-specific or unclassified were then classified into five categories of
492 putative AMGs based on the presence of viral hallmark or non-hallmarks up- or down-stream as
493 mentioned previously[30]. The AMGs with both virus-specific genes ('VV-1' or 'V*-0') were
494 retained for the following analysis. To improve the functional annotations of the putative AMGs
495 and highlight the viral potentials of metabolizing carbohydrates and glycoconjugates, the AMG
496 proteins were also annotated by Carbohydrate-Active enZYmes (CAZY) Database and Kyoto
497 Encyclopedia of Genes and Genomes (KEGG) database using the default settings in addition to
498 the functional annotation databases implemented in geNomad. The putative AMG was assigned
499 to the functional annotation with the highest bitscore (e.g,. duplicate annotations were not
500 allowed). Following Hurwitz and U'Ren[66] and Hurwitz et al[89], we further screened putative
501 AMGs to remove genes not found in KEGG pathways. Additionally, in recognition of the
502 ambiguity in distinguishing genes encoding auxiliary metabolic functions versus core metabolic
503 processes[66], we discuss the resulting set of genes presented here as 'putative AMGs'.

504

505 *Statistical Analysis.*

506

507 All statistical analyses and data visualizations were performed using *R* v 4.1.0[90]. We used
508 the following packages for data manipulation and visualization: ggplot2[91], reshape2[92],
509 pheatmap[93], Hmisc[94], ggpubr[95], RColorBrewer[96], maps[97], stats[98] geosphere[99], plyr[100], dplyr[101],
510 and stringr[102]. Packages pertaining to specific analyses are listed below.
511 We generated rarefaction curves for individual samples and for cumulative sequencing
512 depth (Fig. S1 and S2) using the 'phyloseq' package [103] and custom R plots, respectively.
513 Samples containing less than five vOTUs, viral clusters, or viral Pfams; or less than 100
514 CRISPR-spacer-based host taxon, microbial Pfams, or microbial taxa were removed for visual
515 clarity. Removed samples followed the same general trends as presented in Fig. S1. To visualize
516 saturation across cumulative sequencing depth (Fig. S2), we ordered samples from lowest to
517 highest total assembled bp and progressively added them along the x-axis. On the y-axis, we plot
518 the associated cumulative number of unique attributes.
519 A phylogenetic tree of CRISPR-spaced-based host taxon was generated at the order-level
520 using phyloT v2[104], an online tree generator based on the Genome Taxonomy Database. Then,
521 we visualized the tree in R using the packages 'ggtree'[105], 'treeio'[106], and 'ggnewscale'[107]. To
522 examine relationships between common microbial hosts of soil viruses and soil properties, we
523 first downloaded data describing bulk density, cation exchange capacity, nitrogen, pH, soil
524 organic carbon, and clay content from the SoilGrids250m database[33] using the 'soilDB'
525 package[108]. Mean values of soil properties from 0-5 cm were correlated to the total number of
526 CRISPR spacer hits per microbial ordered using Spearman correlation.
527 Finally, we mapped genes detected across the entire soil virosphere (i.e., all samples
528 combined) to their corresponding KEGG pathways using the 'pathview' package in R[109]. Gene
529 abundances were converted by log base 10 for visualization.

530

531 **DATA AVAILABILITY.**

532 The GSV Atlas is available for download at: https://doi.org/10.25584/2229733. It includes all
533 UViGs regardless of quality (File 1, 616,935 UViGs), a file containing data associated with each
534 contig that passed QA/QC (File 2, 49,649 contigs), predicted viral protein sequences (File 3,
535 402,882 predicted protein sequences), a file containing data associated with each gene (File 4,
536 1,432,147 genes), geographic and physicochemical data of the curated soil samples (File 5, 2,953
537 samples), and a readme file (File 6).

538

539 **CODE AVAILABILITY.** Code for sequence processing is described in Nayafch et al.[35] and is
540 available in the materials associated with those publications. All other code is available upon
541 request.

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. *BioRxiv* (2022) doi:10.1101/2022.03.24.485562.
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).

566 11. Braga, L. P. P. *et al.* Impact of phages on soil bacterial communities and nitrogen
567 availability under different assembly scenarios. *Microbiome* **8**, 52 (2020).
568 12. Wu, R. *et al.* Moisture modulates soil reservoirs of active DNA and RNA viruses.
569 *Commun. Biol.* **4**, 992 (2021).
570 13. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K.
571 Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon
572 cycling in soil. *Proc Natl Acad Sci USA* **116**, 25900–25908 (2019).
573 14. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev.*
574 *Microbiol.* (2022) doi:10.1038/s41579-022-00811-z.
575 15. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat.*
576 *Microbiol.* **3**, 870–880 (2018).
577 16. Suttle, C. A. Marine viruses--major players in the global ecosystem. *Nat. Rev. Microbiol.*
578 **5**, 801–812 (2007).
579 17. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat
580 soil. *Microbiome* **9**, 208 (2021).
581 18. Lee, S. *et al.* Methane-derived carbon flows into host-virus networks at different trophic
582 levels in soil. *Proc Natl Acad Sci USA* **118**, (2021).
583 19. Santos-Medellin, C. *et al.* Viromes outperform total metagenomes in revealing the
584 spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **15**, 1956–1970
585 (2021).
586 20. Snyder, J. C. *et al.* Virus movement maintains local virus population diversity. *Proc Natl*
587 *Acad Sci USA* **104**, 19102–19107 (2007).
588 21. Hesse, U. *et al.* Virome assembly and annotation: A surprise in the namib desert. *Front.*

589 *Microbiol.* **8**, 13 (2017).
590 22. Jin, M. *et al.* Diversities and potential biogeochemical impacts of mangrove soil viruses.
591 *Microbiome* **7**, 58 (2019).
592 23. Trubl, G. *et al.* Soil viruses are underexplored players in ecosystem carbon processing.
593 *mSystems* **3**, (2018).
594 24. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant
595 ocean viruses. *Nature* **537**, 689–693 (2016).
596 25. Sokol, N. W. *et al.* Life and death in the soil microbiome: how ecological processes
597 influence biogeochemistry. *Nat. Rev. Microbiol.* **20**, 415–430 (2022).
598 26. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the
599 marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
600 27. Liang, X. *et al.* Lysogenic reproductive strategies of viral communities vary with soil
601 depth and are correlated with bacterial diversity. *Soil Biol. Biochem* **144**, 107767 (2020).
602 28. Williamson, K. E., Radosevich, M., Smith, D. W. & Wommack, K. E. Incidence of
603 lysogeny within temperate and extreme soil environments. *Environ. Microbiol.* **9**, 2563–
604 2574 (2007).
605 29. Wu, R. *et al.* Structural characterization of a soil viral auxiliary metabolic gene product - a
606 functional chitosanase. *Nat. Commun.* **13**, 5485 (2022).
607 30. Wu, R. *et al.* DNA Viral Diversity, Abundance, and Functional Potential Vary across
608 Grassland Soils with a Range of Historical Moisture Regimes. *MBio* **12**, e0259521 (2021).
609 31. Liang, X. *et al.* Viral abundance and diversity vary with depth in a southeastern United
610 States agricultural ultisol. *Soil Biol. Biochem* **137**, 107546 (2019).
611 32. Jansson, J. K. Soil viruses: Understudied agents of soil ecology. *Environ. Microbiol.* **25**,

612 143–146 (2023).
613 33. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine
614 learning. *PLoS ONE* **12**, e0169748 (2017).
615 34. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG).
616 *Nat. Biotechnol.* **37**, 29–37 (2019).
617 35. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human
618 gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
619 36. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for
620 interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
621 37. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z.
622 Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers
623 during the infection process. *Appl. Microbiol. Biotechnol.* **101**, 3103–3119 (2017).
624 38. Kintz, E. *et al.* A BTP1 prophage gene present in invasive non-typhoidal Salmonella
625 determines composition and length of the O-antigen of the lipopolysaccharide. *Mol.*
626 *Microbiol.* **96**, 263–275 (2015).
627 39. Carini, P. *et al.* Relic DNA is abundant in soil and obscures estimates of soil microbial
628 diversity. *Nat. Microbiol.* **2**, 16242 (2016).
629 40. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial
630 diversity. *Nature* **551**, 457–463 (2017).
631 41. Locey, K. J. *et al.* Dormancy dampens the microbial distance-decay relationship. *Philos.*
632 *Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190243 (2020).
633 42. CAZy - GT4. http://www.cazy.org/GT4.html.
634 43. CAZy - GH73. http://www.cazy.org/GH73.html.

635 44. CAZy - CBM50. http://www.cazy.org/CBM50.html.
636 45. Wommack, K. E. & Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems.
637 *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
638 46. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science* **332**,
639 106–109 (2011).
640 47. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and
641 evolutionary determinants of bacteriophage host range. *Trends Microbiol.* **27**, 51–63
642 (2019).
643 48. Heineman, R. H., Springman, R. & Bull, J. J. Optimal foraging by bacteriophages through
644 host avoidance. *Am. Nat.* **171**, E149-57 (2008).
645 49. Holtzman, T. *et al.* A continuous evolution system for contracting the host range of
646 bacteriophage T7. *Sci. Rep.* **10**, 307 (2020).
647 50. Sant, D. G., Woods, L. C., Barr, J. J. & McDonald, M. J. Host diversity slows
648 bacteriophage adaptation by selecting generalists over specialists. *Nat. Ecol. Evol.* **5**, 350–
649 359 (2021).
650 51. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
651 52. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and its
652 constraints on virus evolution. *Nat. Rev. Microbiol.* **17**, 321–328 (2019).
653 53. Srinivasiah, S. *et al.* Phages across the biosphere: contrasts of viruses in soil and aquatic
654 environments. *Res. Microbiol.* **159**, 349–357 (2008).
655 54. Srinivasiah, S. *et al.* Dynamics of autochthonous soil viral communities parallels
656 dynamics of host communities under nutrient stimulation. *FEMS Microbiol. Ecol.* **91**,
657 (2015).

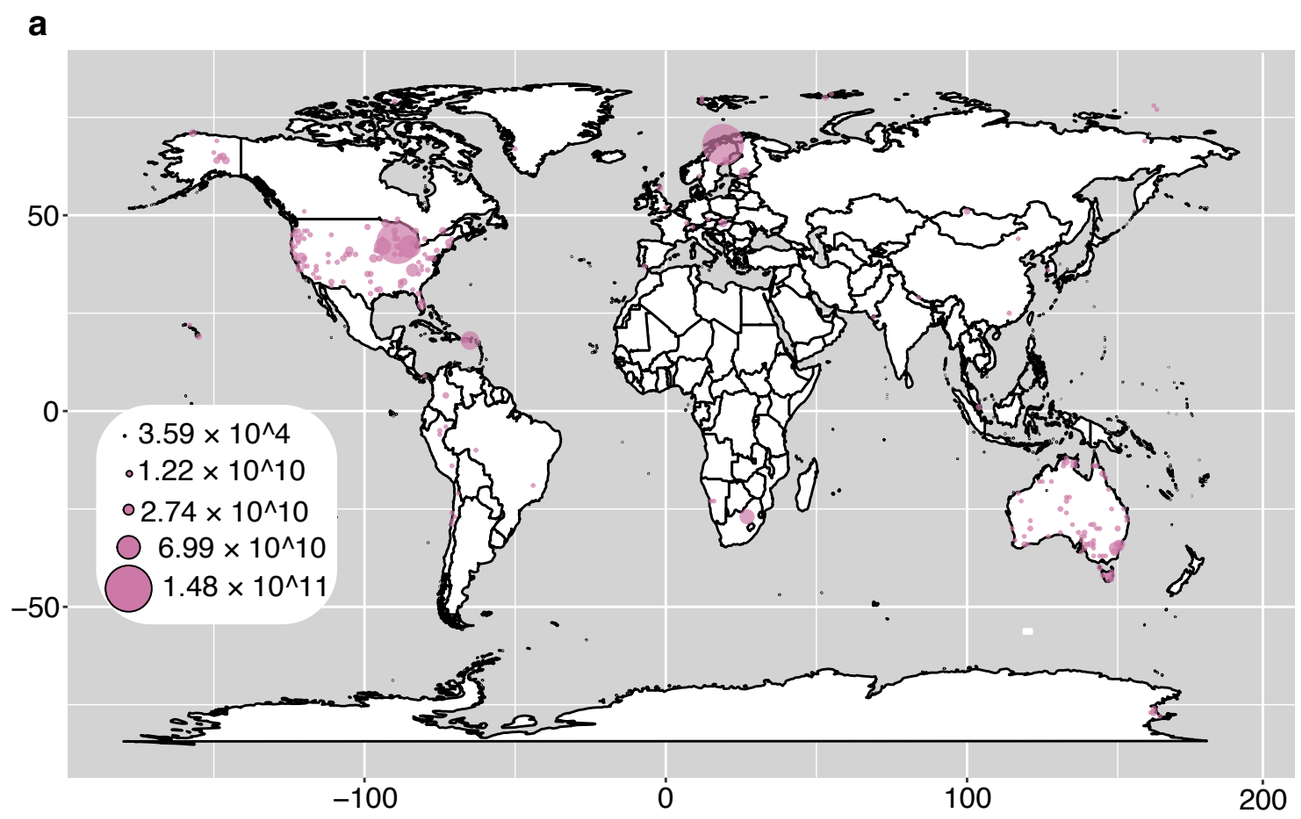
658 55. Romine, M. F. *et al.* Elucidation of roles for vitamin B12 in regulation of folate,
659 ubiquinone, and methionine metabolism. *Proc Natl Acad Sci USA* **114**, E1205–E1214
660 (2017).
661 56. Mattenberger, Y., Mattson, S., Métrailler, J., Silva, F. & Belin, D. 55.1, a gene of
662 unknown function of phage T4, impacts on Escherichia coli folate metabolism and blocks
663 DNA repair by the NER. *Mol. Microbiol.* **82**, 1406–1421 (2011).
664 57. Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L. & Davidson, A. R. The phage
665 lambda major tail protein structure reveals a common evolution for long-tailed phages and
666 the type VI bacterial secretion system. *Proc Natl Acad Sci USA* **106**, 4160–4165 (2009).
667 58. Sgro, G. G. *et al.* Bacteria-Killing Type IV Secretion Systems. *Front. Microbiol.* **10**, 1078
668 (2019).
669 59. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a
670 DNA substrate. *Science* **304**, 1170–1173 (2004).
671 60. Lasker, K., Mann, T. H. & Shapiro, L. An intracellular compass spatially coordinates cell
672 cycle modules in Caulobacter crescentus. *Curr. Opin. Microbiol.* **33**, 131–139 (2016).
673 61. Tan, Y. *et al.* Organic fertilizers shape soil microbial communities and increase soil amino
674 acid metabolites content in a blueberry orchard. *Microb. Ecol.* **85**, 232–246 (2023).
675 62. Wang, D., Chadwick, D. R., Hill, P. W., Ge, T. & Jones, D. L. Rapid microbial uptake and
676 mineralization of 14C-labelled cysteine and methionine along a grassland productivity
677 gradient. *Soil Biol. Biochem* **180**, 109022 (2023).
678 63. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the
679 dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**,
680 892–899 (2007).

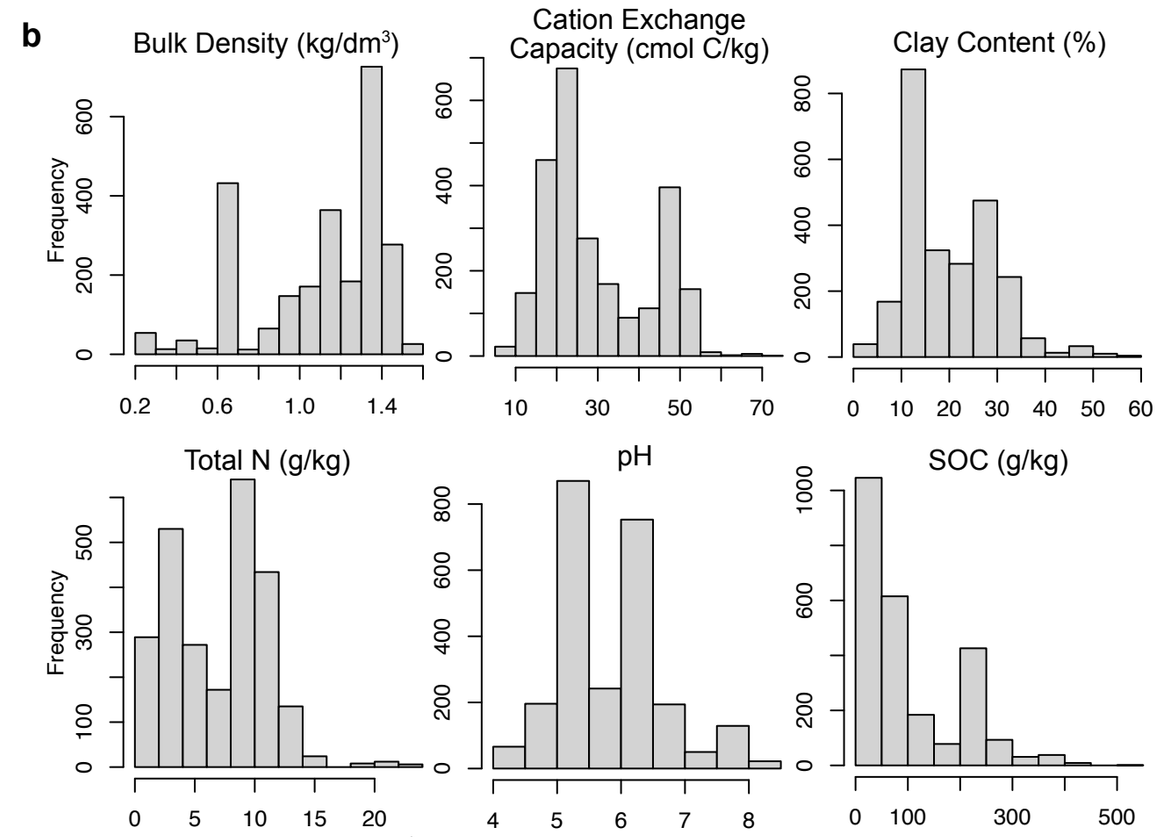
681 64. Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. *Annu. Rev. Genet.* **42**,
682 647–681 (2008).
683 65. Kühlbrandt, W. Structure and Mechanisms of F-Type ATP Synthases. *Annu. Rev.*
684 *Biochem.* **88**, 515–549 (2019).
685 66. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems.
686 *Curr. Opin. Microbiol.* **31**, 161–168 (2016).
687 67. Daniel, R. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).
688 68. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T.
689 Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems*
690 **3**, (2018).
691 69. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil.
692 *Methods Mol. Biol.* **501**, 15–21 (2009).
693 70. Trubl, G. *et al.* Towards optimized viral metagenomes for double-stranded and single-
694 stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
695 71. Zablocki, O. *et al.* VirION2: a short- and long-read sequencing and informatics workflow
696 to study the genomic diversity of viruses in nature. *PeerJ* **9**, e11088 (2021).
697 72. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and
698 abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
699 73. Guerra, C. A. *et al.* Blind spots in global soil biodiversity and ecosystem function
700 research. *Nat. Commun.* **11**, 3870 (2020).
701 74. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.8: overview and updates.
702 *Nucleic Acids Res.* **49**, D723–D733 (2021).
703 75. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative

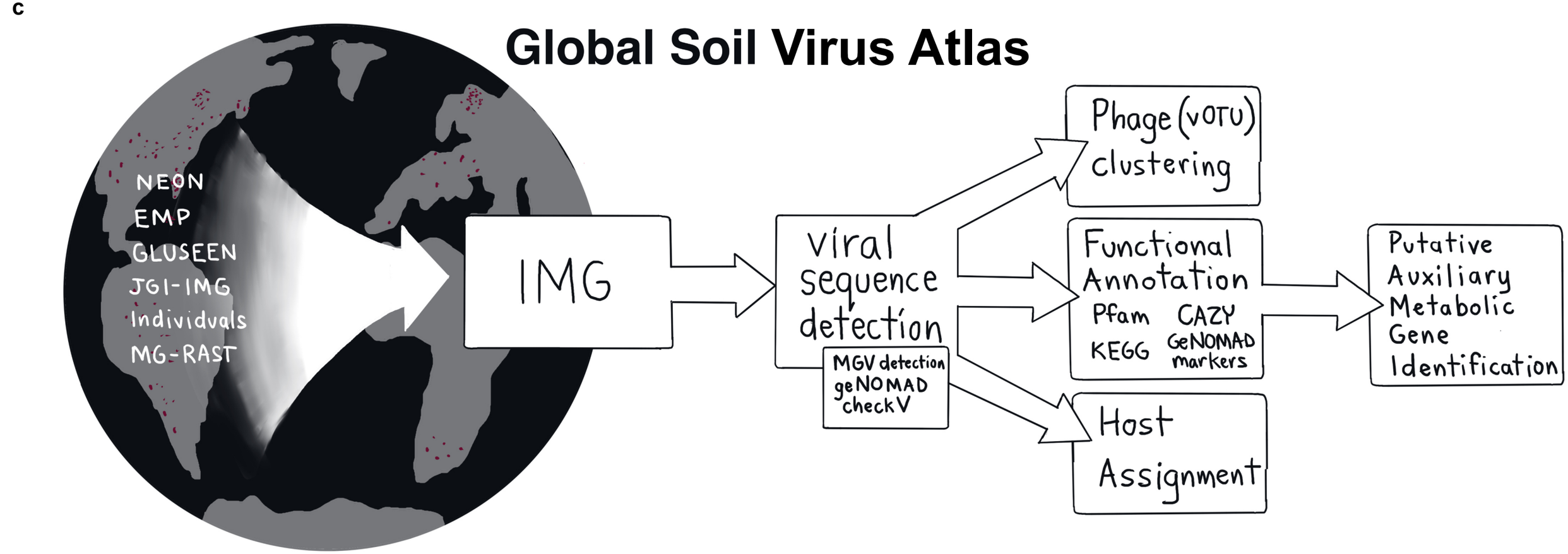
704 analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–
705 D677 (2019).
706 76. Clum, A. *et al.* DOE JGI metagenome workflow. *mSystems* **6**, (2021).
707 77. Nelson, W. C. *et al.* Terabase metagenome sequencing of grassland soil microbiomes.
708 *Microbiol. Resour. Announc.* **9**, (2020).
709 78. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**,
710 D427–D432 (2019).
711 79. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer
712 based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*
713 **5**, 69 (2017).
714 80. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat.*
715 *Biotechnol.* (2023) doi:10.1038/s41587-023-01953-y.
716 81. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-
717 assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
718 82. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
719 using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
720 83. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM. J. Matrix Anal.*
721 *& Appl.* **30**, 121–141 (2008).
722 84. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–
723 509 (2021).
724 85. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
725 microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
726 86. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over

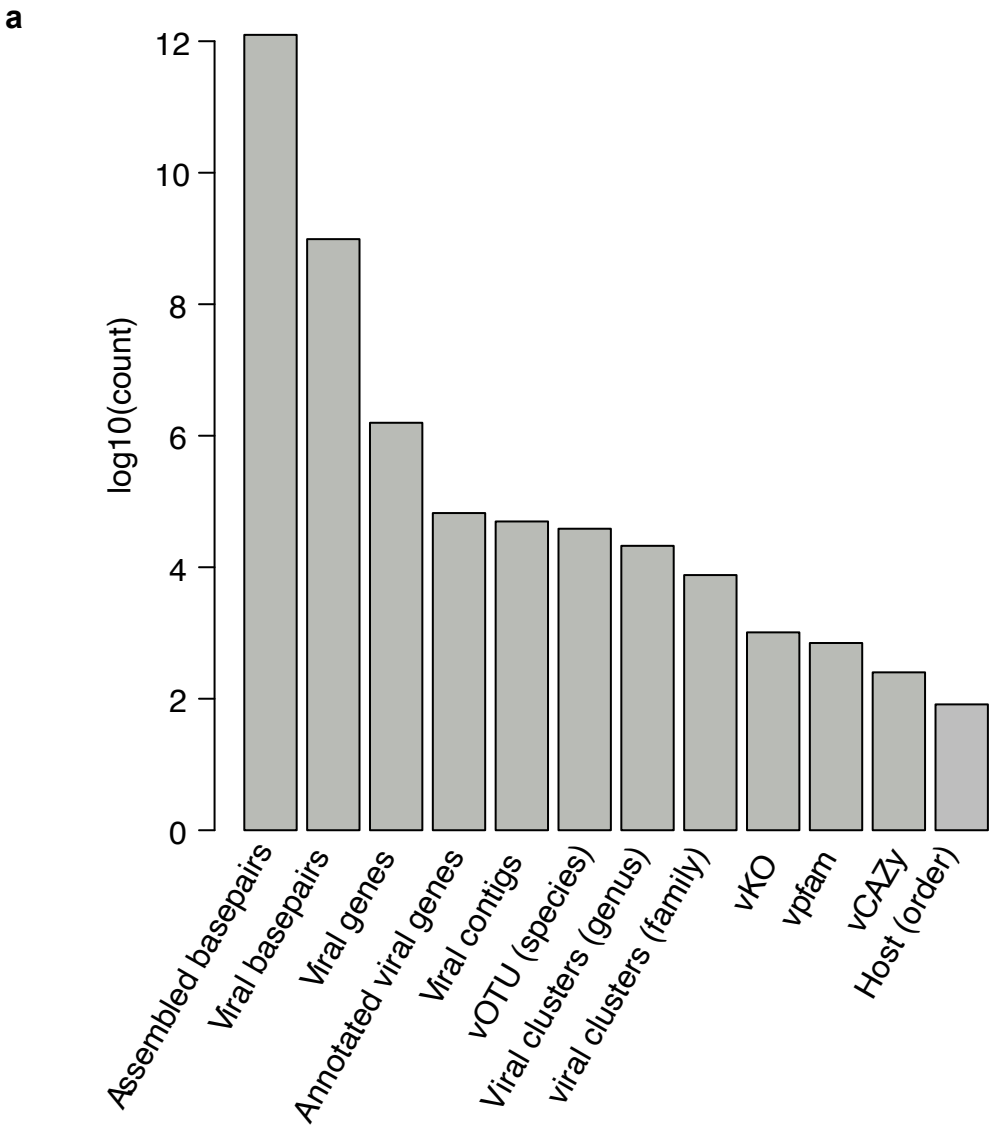
727 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**,
728 649-662.e20 (2019).
729 87. Merrill, B. D. *et al.* Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers
730 Vanishing Gut Microbes. *BioRxiv* (2022) doi:10.1101/2022.03.30.486478.
731 88. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes
732 within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic*
733 *Acids Res.* **51**, D733–D743 (2023).
734 89. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic
735 niche specialization in the "core" and "flexible" Pacific Ocean Virome. *ISME J.* **9**, 472–
736 484 (2015).
737 90. R Core Team. *R: A language and environment for statistical computing*. (R Foundation
738 for Statistical Computing, 2021).
739 91. Wickham, H. & Chang, W. Package "ggplot2." *Create elegant data …* (2016).
740 92. Wickham, H. reshape2: flexibly reshape data: a reboot of the reshape package.
741 93. Kolde, R. & Kolde, M. R. Package "pheatmap." *R Package* (2018).
742 94. Jr, F. H. & Jr, M. H. Package "hmisc." *CRAN2018* (2019).
743 95. Goslee, S. C. & Urban, D. L. The ecodist Package for Dissimilarity-based Analysis of
744 Ecological Data. *J. Stat. Softw.* **22**, (2007).
745 96. Neuwirth, E. & Brewer, R. C. ColorBrewer palettes. *R package version* (2014).
746 97. Becker, R. A., Wilks, A. R., Brownrigg, R. & Minka, T. P. maps: Draw geographical
747 maps.
748 98. R: The R Stats Package. https://stat.ethz.ch/R-manual/R-
749 devel/library/stats/html/00Index.html.

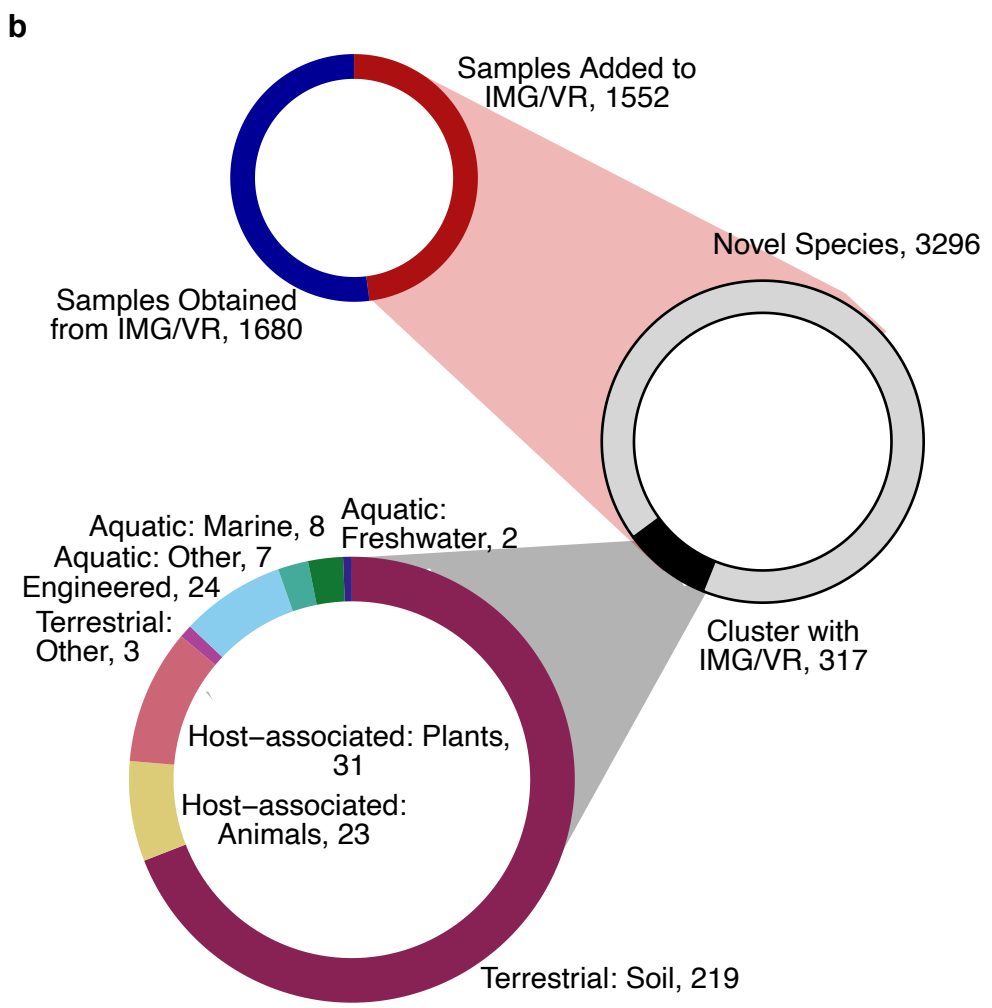
750 99. Hijmans, R. J., Williams, E., Vennes, C. & Hijmans, M. R. J. Package "geosphere."
751 100. Wickham, H. Tools for Splitting, Applying and Combining Data [R package plyr version
752 1.8.8]. (2022).
753 101. A Grammar of Data Manipulation [R package dplyr version 1.1.2]. https://cran.r-
754 project.org/web/packages/dplyr/index.html (2023).
755 102. Wickham, H. Simple, Consistent Wrappers for Common String Operations [R package
756 stringr version 1.5.0]. (2022).
757 103. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive
758 analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
759 104. phyloT : a phylogenetic tree generator. https://phylot.biobyte.de/.
760 105. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: An R package for
761 visualization and annotation of phylogenetic trees with their covariates and other
762 associated data. *Methods Ecol. Evol.* **8**, 28–36 (2016).
763 106. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with
764 Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
765 107. Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. *R package version*
766 *0.4* (2020).
767 108. DE Beaudette, Skovlin, J., Roecker, S. & Beaudette, M. D. E. Package "soilDB." (2022).
768 109. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data
769 integration and visualization. *Bioinformatics* **29**, 1830–1831 (2013).

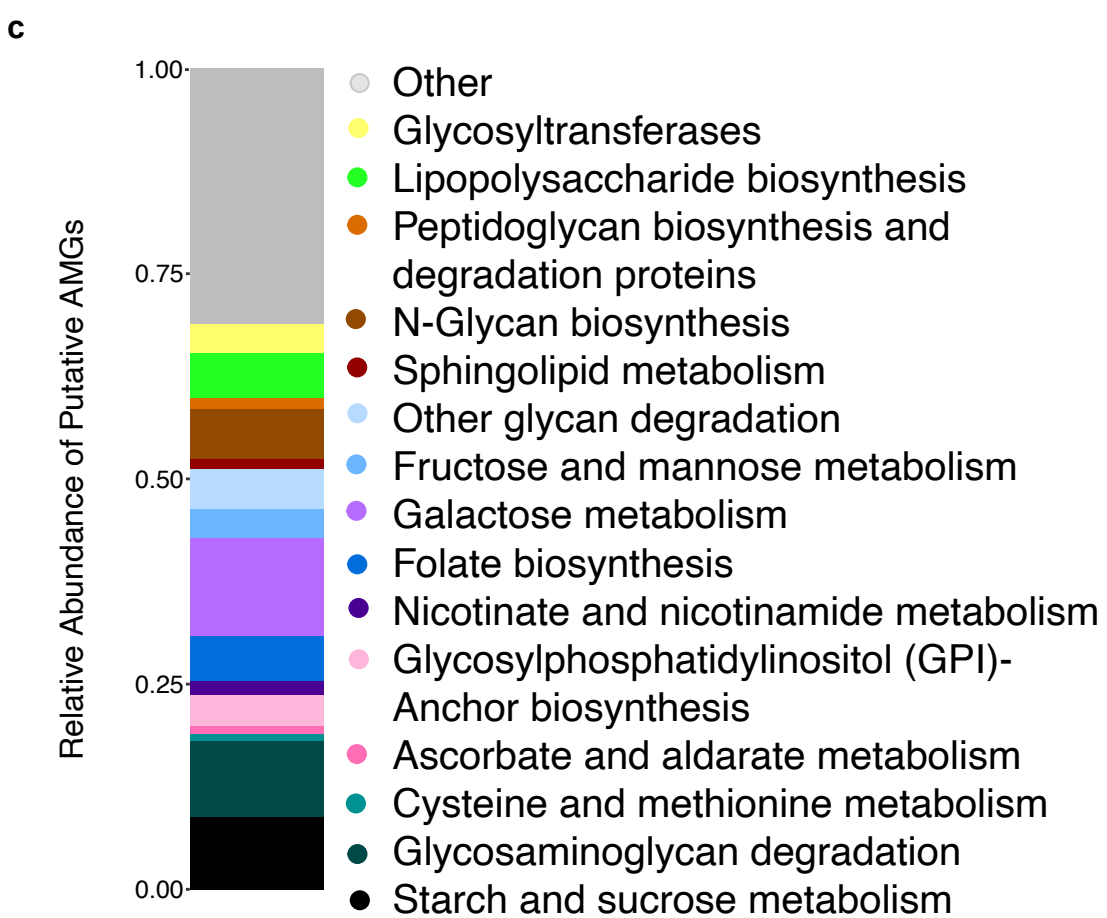

a
50
0
−50
−100
0
100
200
3.59 × 10^4
1.22 × 10^10
2.74 × 10^10
6.99 × 10^10
1.48 × 10^11

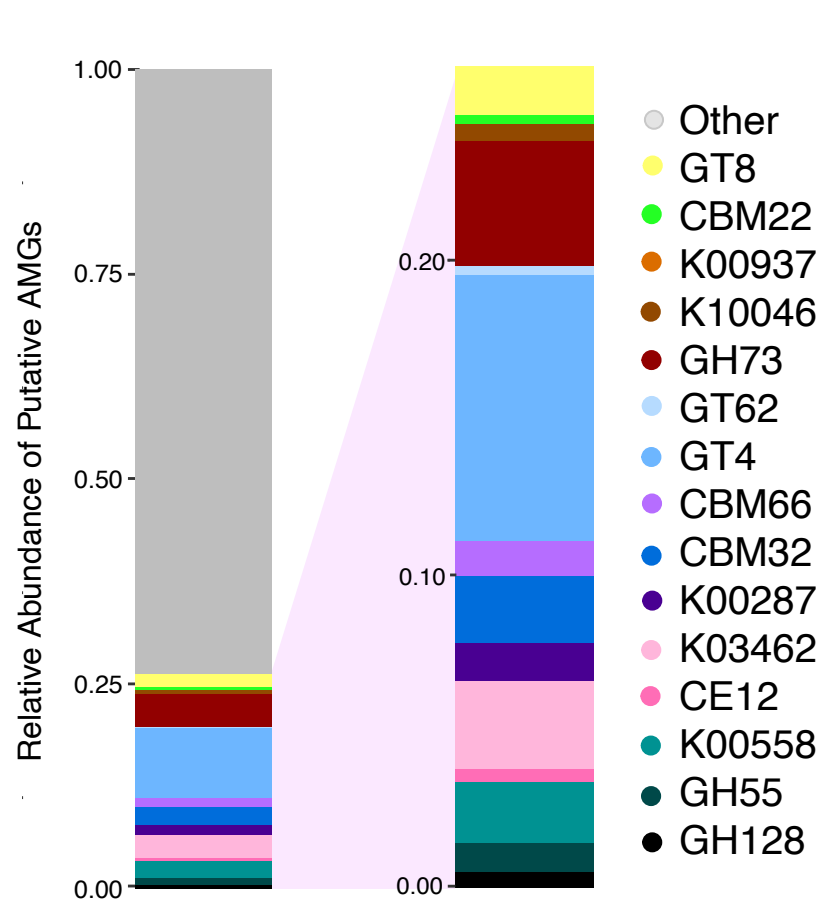

b
Bulk Density (kg/dm³)
Cation Exchange Capacity (cmol C/kg)
Clay Content (%)
Total N (g/kg)
pH
SOC (g/kg)
Frequency

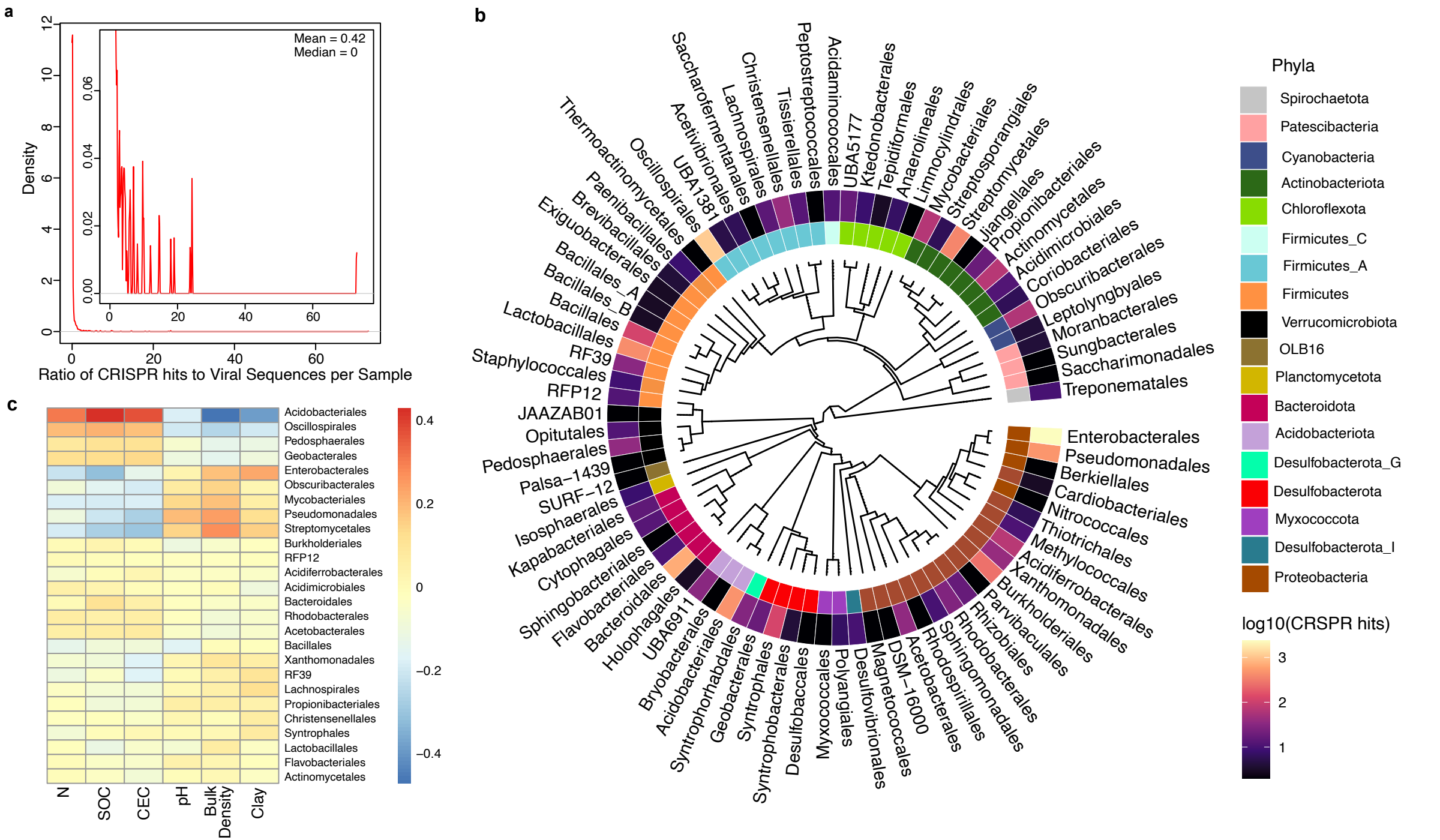

c
Global Soil Virus Atlas
NEON
EMP
GLUSEEN
JGI-IMG
Individuals
MG-RAST
IMG
Viral sequence detection
MGV detection
geNOMAD
checkV
Phage (vOTU) clustering
Functional Annotation
Pfam
CAZY
KEGG
GeNOMAD markers
Host Assignment
Putative Auxiliary Metabolic Gene Identification

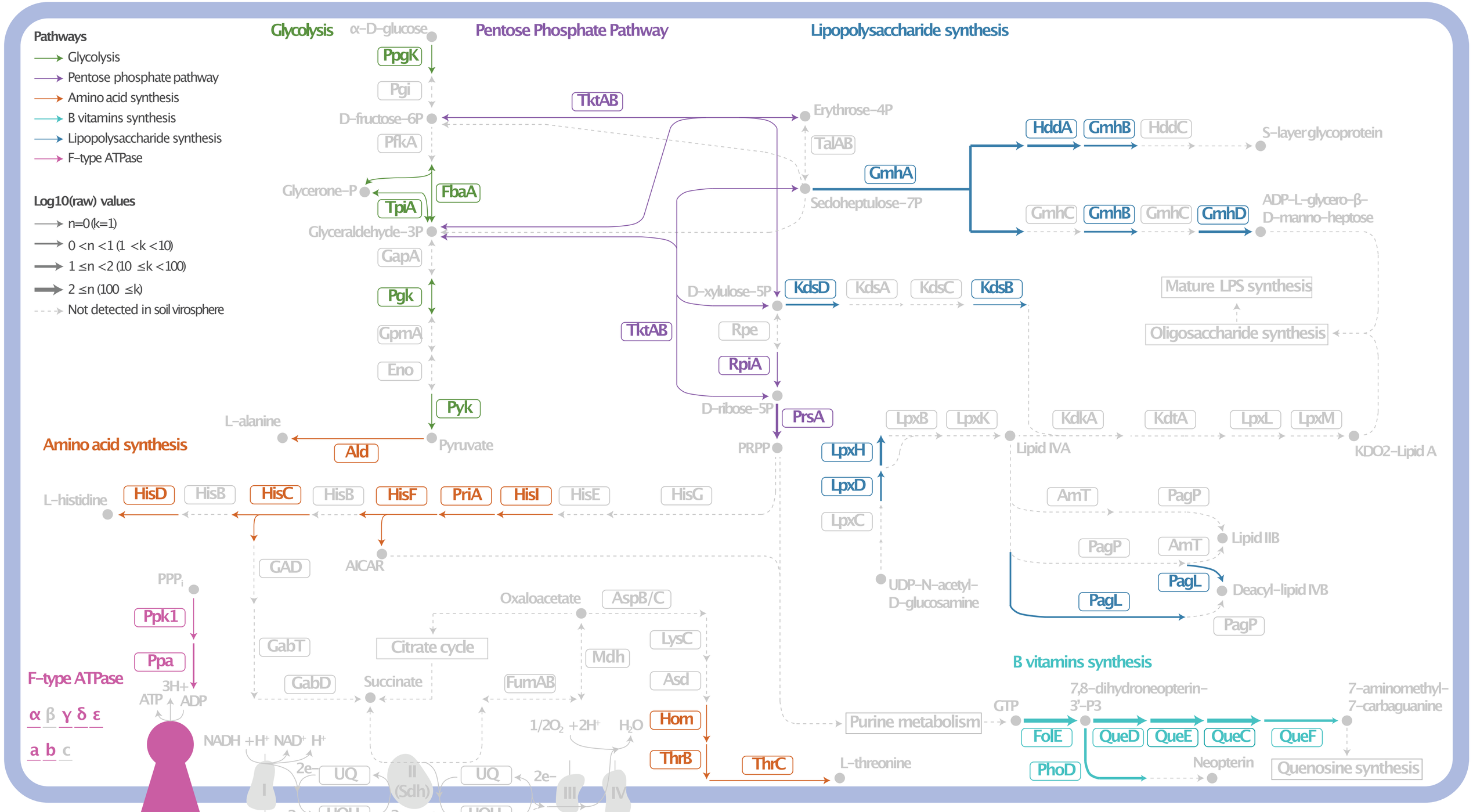

a
log10(count)
0
2
4
6
8
10
12
Assembled basepairs
Viral basepairs
Viral genes
Annotated viral genes
Viral contigs
vOTU (species)
Viral clusters (genus)
viral clusters (family)
vKO
vpfam
vCAZy
Host (order)


b
Samples Added to IMG/VR, 1552
Samples Obtained from IMG/VR, 1680
Novel Species, 3296
Cluster with IMG/VR, 317
Aquatic: Marine, 8
Aquatic: Freshwater, 2
Aquatic: Other, 7
Engineered, 24
Terrestrial: Other, 3
Host−associated: Plants, 31
Host−associated: Animals, 23
Terrestrial: Soil, 219


c
Relative Abundance of Putative AMGs
0.00
0.25
0.50
0.75
1.00
Other
Glycosyltransferases
Lipopolysaccharide biosynthesis
Peptidoglycan biosynthesis and degradation proteins
N-Glycan biosynthesis
Sphingolipid metabolism
Other glycan degradation
Fructose and mannose metabolism
Galactose metabolism
Folate biosynthesis
Nicotinate and nicotinamide metabolism
Glycosylphosphatidylinositol (GPI)-Anchor biosynthesis
Ascorbate and aldarate metabolism
Cysteine and methionine metabolism
Glycosaminoglycan degradation
Starch and sucrose metabolism


d
Relative Abundance of Putative AMGs
0.00
0.25
0.50
0.75
1.00
0.00
0.10
0.20
Other
GT8
CBM22
K00937
K10046
GH73
GT62
GT4
CBM66
CBM32
K00287
K03462
CE12
K00558
GH55
GH128


a
Mean = 0.42
Median = 0
Density
Ratio of CRISPR hits to Viral Sequences per Sample
b
Phyla
Spirochaetota
Patescibacteria
Cyanobacteria
Actinobacteriota
Chloroflexota
Firmicutes_C
Firmicutes_A
Firmicutes
Verrucomicrobiota
OLB16
Planctomycetota
Bacteroidota
Acidobacteriota
Desulfobacterota_G
Desulfobacterota
Myxococcota
Desulfobacterota_I
Proteobacteria
log10(CRSPR hits)
Treponematales
Saccharimonadales
Sungbacterales
Moranbacterales
Leptolyngbyales
Obscuribacterales
Coriobacteriales
Acidimicrobiales
Actinomycetales
Propionibacteriales
Jiangellales
Streptomycetales
Streptosporangiales
Mycobacteriales
Limnocylindrales
Anaerolineales
Tepidiformales
Ktedonobacterales
UBA5177
Acidaminococcales
Peptostreptococcales
Tissierellales
Christensenellales
Lachnospirales
Saccharofermentanales
Acetivibrionales
UBA1381
Oscillospirales
Thermoactinomycetales
Paenibacillales
Brevibacillales
Exiguobacterales
Bacillales_A
Bacillales_B
Bacillales
Lactobacillales
RF39
Staphylococcales
RFP12
JAAZAB01
Opitutales
Pedosphaerales
Palsa−1439
SURF−12
Isosphaerales
Kapabacteriales
Cytophagales
Sphingobacteriales
Flavobacteriales
Bacteroidales
Holophagales
UBA6911
Bryobacterales
Acidobacteriales
Syntrophorhabdales
Geobacterales
Syntrophales
Syntrophobacterales
Desulfobaccales
Myxococcales
Polyangiales
Desulfovibrionales
Magnetococcales
DSM−16000
Acetobacterales
Rhodospirillales
Sphingomonadales
Rhodobacterales
Rhizobiales
Parvibaculales
Burkholderiales
Xanthomonadales
Acidiferrobacterales
Methylococcales
Thiotrichales
Nitrococcales
Cardiobacteriales
Berkiellales
Pseudomonadales
Enterobacterales
c
Acidobacteriales
Oscillospirales
Pedosphaerales
Geobacterales
Enterobacterales
Obscuribacterales
Mycobacteriales
Pseudomonadales
Streptomycetales
Burkholderiales
RFP12
Acidiferrobacterales
Acidimicrobiales
Bacteroidales
Rhodobacterales
Acetobacterales
Bacillales
Xanthomonadales
RF39
Lachnospirales
Propionibacteriales
Christensenellales
Syntrophales
Lactobacillales
Flavobacteriales
Actinomycetales
N
SOC
CEC
pH
Bulk Density
Clay


Pathways
Glycolysis
Pentose phosphate pathway
Amino acid synthesis
B vitamins synthesis
Lipopolysaccharide synthesis
F-type ATPase
Log10(raw) values
n=0 (k=1)
0 <n <1 (1 <k <10)
1 ≤n <2 (10 ≤k <100)
2 ≤n (100 ≤k)
Not detected in soil virosphere
Glycolysis
α-D-glucose
PpgK
Pgi
D-fructose-6P
PfkA
Glycerone-P
TpiA
FbaA
Glyceraldehyde-3P
GapA
Pgk
GpmA
Eno
Pyk
Pyruvate
Ald
L-alanine
Pentose Phosphate Pathway
TktAB
Erythrose-4P
TalAB
Sedoheptulose-7P
D-xylulose-5P
Rpe
RpiA
D-ribose-5P
PrsA
PRPP
Lipopolysaccharide synthesis
GmhA
HddA
GmhB
HddC
S-layer glycoprotein
GmhC
GmhB
GmhC
GmhD
ADP-L-glycero-β-D-manno-heptose
KdsD
KdsA
KdsC
KdsB
Mature LPS synthesis
Oligosaccharide synthesis
LpxB
LpxK
KdkA
KdtA
LpxL
LpxM
Lipid IVA
KDO2-Lipid A
LpxH
LpxD
LpxC
UDP-N-acetyl-D-glucosamine
AmT
PagP
Lipid IIB
PagL
Deacyl-lipid IVB
Amino acid synthesis
L-histidine
HisD
HisB
HisC
HisB
HisF
PriA
HisI
HisE
HisG
GAD
AICAR
F-type ATPase
PPPi
Ppk1
Ppa
3H+
ATP
ADP
α β γ δ ε
a b c
Oxaloacetate
AspB/C
GabT
Citrate cycle
Mdh
LysC
GabD
Succinate
FumAB
Asd
Hom
ThrB
ThrC
L-threonine
Purine metabolism
B vitamins synthesis
GTP
7,8-dihydroneopterin-3'-P3
7-aminomethyl-7-carbaguanine
FolE
QueD
QueE
QueC
QueF
PhoD
Neopterin
Quenosine synthesis
NADH +H+ NAD+ H+
2e-
UQ
UQH2
I
II (Sdh)
III
IV
4H+
1/2O2 +2H+
H2O
2 x 2H+ 2H+