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

Emily B. Graham[1,2*], Antonio Pedro Camargo[3], Ruonan Wu[1], Russell Y. Neches[3,4], Matt Nolan[3], David Paez-Espino[3], Nikos C. Kyrpides[3], Janet K. Jansson[1], Jason E. McDermott[1,5], Kirsten S. Hofmockel[1,6], 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 compiled from 2,953 previously sequenced soil metagenomes and 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 an extensive characterization of soil viral diversity and provides a foundation for developing testable hypotheses regarding the role of the virosphere in the soil microbiome and global biogeochemistry.

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

High soil viral diversity may be due to the structural and/or physico-chemical heterogeneity of 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

77 *Global Soil Virus (GSV) Atlas.*

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

89 We also collected associated environmental parameters describing each sample from the
90 SoilGrids250m database[33]. We assayed a wide variety of soils that ranged from bulk density of
91 0.24-1.56 (kg/dm^3), cation exchange capacity of 6.8-71 (CEC; cmol(c)/kg), nitrogen content of
92 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
93 2.7-57.1 (%) (Fig. 1).

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

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

102 We identified 1,432,147 viral genes, of which only 260,258 (~18%) were found in at
103 least one annotation database (1,022, 3,634, and 145 unique KO, Pfam, and CAZy annotations,
104 Fig. 2). After filtering to putative AMGs (see methods)[30], we found 5,043 genes that mapped to
105 83 KEGG pathways (1,941 KO and 3,357 CAZyme, some genes were associated with both KO
106 and CAZyme). The median per sample putative AMG abundance (gene copies per sample) was
107 ~19 genes (median = 4, reflecting skewing from a few samples with high AMG abundance).

108 Some KEGG pathways represented by the most putative AMGs were associated with
109 major soil carbon cycling processes (Galactose metabolism and Starch and sucrose metabolism).
110 Likewise, at the level of gene annotations, the most common putative AMGs suggested a role for
111 viruses in soil carbon cycles; including Glycosyltransferases 4 (GT4), glycosylhydrolases 73
112 (GH73), and Carbohydrate Binding Module 50 (CBM50). Other abundant KEGG pathways and
113 gene annotations (Fig. 2c-d) included Glycosaminoglycan degradation, N-Glycan biosynthesis,
114 Folate biosynthesis, 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase (K01737),
115 and 7-carboxy-7-deazaguanine synthase (K10026).

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

123 Joint Genome Institute (JGI) recommended sequencing depth for soil samples (45 GB). When
124 considering cumulative unique attributes versus sequencing depth (Fig. S2), relationships in
125 vOTUs and viral clusters displayed slight curvature, while viral Pfams neared saturation.
126
127 *Microbial Hosts of Soil Viruses.*
128 We connected 1,450 viruses to putative hosts of 82 different bacterial and archaeal orders
129 with CRISPR spacers. This equates to 2.78% of QA/QC-ed viral contigs that were associated
130 with CRISPR spacer hits; roughly 70% more host assignments than in another recent
131 assessment[4,36]. While we observed a maximum of 73 vOTUs per host (i.e., CRISPR spacer), the
132 mean overall vOTU per host ratio was 0.42 (median = 0), reflecting the predominance of unique
133 host associations for individual vOTUs.
134 Out of 1,223 samples with at least one vOTUs assigned to a host, only 72 samples had an
135 average of more than one host sequence per vOTU, underscoring the low abundance of detected
136 hosts across all soils. An average of 0.64 unique host orders were detected per sample, with a
137 maximum ratio of CRISPR spacer hits to viral sequences of 73. Further, samples with a high
138 ratio of vOTU:host almost exclusively were matched to host sequences from a single microbial
139 order, reflecting high phylogenetic conservation of host associations. Of the 10 samples with the
140 highest CRISPR spacer sequence to viral sequence ratio, only 1 contained a CRISPR spacer
141 matching more than one microbial order.
142 The most prevalent host taxa were distributed across distantly-related phyla, including
143 members of prominent soil orders such as Pseudomonadales, Burkholderiales, Acidobacteriales,
144 and Bacteroidales (Fig. 3b). The frequency of CRISPR hits associated with Acidobacteriales,
145 Oscillospirales, Pedosphaerales, and Geobacterales were positively correlated to soil nitrogen,

146 organic carbon, and CEC; while Enterobacterales, Obscuribacterales, Mycobacteriales,
147 Pseudomonadales, and Streptomycetales were positively correlated to soil bulk density, and to a
148 lesser extent, pH and clay.
149
150 *Metabolic potential encoded by the soil virosphere.*

151 Because viral gene annotations were sparsely distributed across many functions, we
152 screened all viral genes (regardless of assignment by the AMG pipeline) against KEGG
153 pathways to understand relationships among genes in the context of known metabolic processes.
154 Across the entire soil virosphere, we uncovered portions of KEGG pathways that were mostly
155 complete, including functions involved in amino acid and sugar metabolism and biosynthesis,
156 antibacterial mechanisms, nucleotide synthesis, and other viral functions (e.g., infection
157 strategies, Extended Data and Fig. 4). For instance, genes associated with DNA mismatch repair
158 (map03430), homologous recombination (map03440), and base excision repair (map03410) were
159 prevalent (Fig. S3). Folate biosynthesis was also common in the soil virosphere (map00670,
160 map00790, Fig. 4). Bacterial secretion systems (map03070, Fig. S4), which may be
161 evolutionarily derived from phages[37], and the *Caulobacter* cell cycle (map04112, Fig. S5), which
162 has a distinct division pattern[38], were rife across soils. The GSV Atlas also contained many viral
163 amino acid biosynthesis/degradation pathways that could be critical in viral life cycles (e.g.,
164 map00250, map00260, map00270, map00330, map00340, Fig. S6). Finally, we found nearly
165 complete portions of energy-generating pathways including the pentose phosphate pathways and
166 F-type ATPase-mediated portions of photosynthesis. Lipopolysaccharide pathway-related genes
167 (LPS) that may be important as host receptors for bacteriophage and prevention of superinfection
168 were also prevelant[39,40].

169 **DISCUSSION**

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

183 Analysis of the GSV Atlas suggests that extreme spatial heterogeneity may be a key
184 feature of the soil virosphere at the global scale. While rapid viral spatial turnover was recently
185 observed across short spatial scales (<10-20m)[8,10], there has been no such demonstration of viral
186 biogeography across the world. We propose that high rates of spatial turnover could result from
187 low dispersal rates or distinct temporal dynamics of viral communities relative to other
188 organisms. For example, while dormant microorganisms and relic DNA can persist for months or
189 more[41–43], the burst of viral cells associated with active infections may generate short-lived
190 pulses of distinct viral communities that do not contribute to relic DNA due to their
191 comparatively small genome sizes versus microorganisms. Additionally, the apparent

192 discrepancies between microbial and viral dispersal processes could be due to the presence of
193 free viruses that are not actively involved in microbial infection[14], smaller viral genomes that
194 could facilitate physical protection, differences in traits that facilitate dispersal between viruses
195 and microorganisms, variation in bioinformatic pipelines, and/or other ecological differences
196 between viruses and microorganisms.

197 Together, these factors make characterizing the soil virosphere a challenge for the
198 coming decade. When examining individual soil samples, the number of new and unique viral
199 attributes (e.g., vOTUs, clusters, Pfams) was linearly related to sequencing depth, suggesting that
200 new viral discoveries are likely to continue with increasingly deep sequencing (Fig. S1). This
201 contrasts with rarefaction curves of the soil microbiome and of microbial hosts of soil viruses,
202 which both asymptoted well before sequencing depths of typical soil microbial investigations.
203 Still, when looking at the cumulative number of unique viral attributes detected in all samples
204 collectively (Fig. S2), many viral attributes began to saturate with sequencing depth. This
205 suggests that while individual samples do not capture soil viral diversity, we can begin to
206 constrain the extent of diversity when sequences from thousands of existing samples are
207 aggregated.

208 Functional diversity encoded by the GSV Atlas revealed the potential for soil viruses to
209 impact biogeochemical cycles; in particular by supporting organic matter decomposition. KEGG
210 pathways and gene annotations represented by the most putative AMGs were related to the
211 metabolism and/or production of sugars common to soils including sucrose, mannose,
212 glucosamine, and maltose[44,45]; as well as the decomposition of chitin – one of the most abundant
213 carbon molecules in soil[46]. Our results are consistent with previous work from single locations
214 that have hinted at a wide range of possible soil viral AMGs, including glycoside hydrolases,

215 carbohydrate esterases, and carbohydrate-binding modules[15,23,31]. Given that a large proportion of
216 soil microorganisms are infected by viruses at any given time[47], AMGs encoded by soil viruses
217 have the potential to impact global biogeochemical cycles[15,22,23,31]. The thousands of putative
218 AMGs identified here represent the most extensive survey to date and further impress the
219 importance of the soil virosphere as a reservoir for biogeochemical potential.

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

233 The ultimate impact of viral predation on soil functions is at least partially associated
234 with the taxonomic distribution of hosts and their variation across soil habitats. Host sequences
235 spanned nearly every major soil microbial clade, consistent with other recent studies (Fig.
236 3)[22,23,31]. This taxonomic breadth suggests a role for the soil virosphere in most soil habitats.
237 Moreover, some hosts were susceptible to changes in the environment, which may reflect

238 environmental filtering on host communities (which in turn, determine the amount and type of
239 viruses present) or on viruses directly which subsequently impacts host community
240 composition[15,27,55,56]. Viral infections have been previously linked to soil parameters including
241 moisture[12,30] and carbon and nitrogen content[9]. In our analysis, bulk density may serve as a proxy
242 for hydrologic connectivity in the soil matrix. For example, low hydrologic connectivity may
243 create 'spatial refuges' for soil bacteria from viral infections[8], influence the virus-host encounter
244 rates, and thus structure the soil virosphere and its hosts. Nutrient amendments are also
245 considered to be drivers of the soil virosphere, supporting the relationship we observed between
246 carbon, nitrogen, and host taxa.

247 When examining the functional potential of the soil virosphere, we detected many
248 hallmarks of viral activity—including genes associated with cell lysis, DNA repair/replication,
249 and other infection signatures—and viral amino acid biosynthesis/degradation pathways that
250 could be critical in viral life cycles (Fig. 4, Fig. S3-S6). The prevalence of viral genes associated
251 with central microbial functions highlights the potential importance of viral activity in soils and
252 the need for targeted approaches to quantify the extent and impact of viral gene expression. For
253 instance, folate and other B vitamins may be logical targets for pathogens as they are key to
254 bacterial growth(map00670, map00790, Fig. 4)[57,58]. Type IV secretion systems can be used by
255 bacteria to secrete toxins[59] or as a method for DNA transfer through membranes.[60] The
256 *Caulobacter* cell cycle (map04112, Fig. S5) is another promising indicator of viral infections due
257 to its distinct cell division process[38]. Finally, amino acids are building blocks for cellular material
258 and also support soil biogeochemical cycles, as they can enhance carbon cycling through priming
259 effects and/or enhanced nutrient availability[61,62] (e.g., map00500, map00052, map00051, Fig. 4).
260 Collectively, these pathways demonstrate several possibilities for soil viral impacts on processes

261 that are central to microbial metabolism and biogeochemical cycling of elements in soil.

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

281 The field of soil viral ecology is poised for rapid expansion; yet several challenges
282 remain in fully characterizing soil viral diversity and function. Overcoming these methodological
283 and ecological hurdles will require broad participation from global researchers. Below, we

284 present a summary of issues, from our perspective, facing the current generation of soil viral
285 ecologists and suggestions for surmounting them.

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

298 Knowledge about soil viral diversity and function is also limited by gaps in field and
299 laboratory experiments. The GSV Atlas demonstrates that extensive, spatially-explicit sampling
300 is needed to capture the high spatial turnover of the soil virosphere. The spatial coverage of most
301 'global' ecological studies, including this one, often suffers from large data gaps[73]. Concerted
302 efforts are needed to sample wide spatial domains, including historically undersampled regions,
303 given the high viral diversity uncovered by the GSV Atlas. Expansion of the known virosphere
304 in this way will also help to facilitate tool development. Although we did not assess temporal
305 dynamics, temporally-explicit approaches are likewise needed to characterize temporal dynamics
306 in soil viral communities. Further, our functional annotation of viral contigs revealed diverse

genes associated with functions relevant to both viral and microbial communities, and it is impossible to know the true functions of viral genes without targeted functional assays. We therefore propose that experiments targeting the expression and auxiliary metabolic function of viral genes are needed to properly assess AMGs in viral communities.

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

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

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

346 **FIGURES**

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

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

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

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

369 phosphate pathway (map00030), and vitamin B- and amino acid-related KEGG pathways in the
370 soil virosphere. Genes detected in the soil virosphere are colored according to pathway type.
371 Undetected genes and associated metabolites (unmeasured) are grayed out. Arrow width denotes
372 gene abundance.

373
374

375 **REFERENCES**

376

377 1. Williamson, K. E., Fuhrmann, J. J., Wommack, K. E. & Radosevich, M. Viruses in soil
378 ecosystems: an unknown quantity within an unexplored territory. Annu. Rev. Virol. 4, 201–219
379 (2017).
380 2. Pratama, A. A. & van Elsas, J. D. The "Neglected" Soil Virome - Potential Role and
381 Impact. Trends Microbiol. 26, 649–662 (2018).
382 3. Kuzyakov, Y. & Mason-Jones, K. Viruses in soil: Nano-scale undead drivers of microbial
383 life, biogeochemical turnover and ecosystem functions. Soil Biol. Biochem 127, 305–317 (2018).
384 4. Roux, S. & Emerson, J. B. Diversity in the soil virosphere: to infinity and beyond?
385 Trends Microbiol. 30, 1025–1035 (2022).
386 5. Kimura, M., Jia, Z.-J., Nakayama, N. & Asakawa, S. Ecology of viruses in soils: Past,
387 present and future perspectives. Soil Sci. Plant Nutr. 54, 1–32 (2008).
388 6. Cobián Güemes, A. G. et al. Viruses as winners in the game of life. Annu. Rev. Virol. 3,
389 197–214 (2016).
390 7. Ter Horst, A. M. et al. Minnesota peat viromes reveal terrestrial and aquatic niche
391 partitioning for local and global viral populations. Microbiome 9, 233 (2021).
392 8. Santos-Medellín, C. et al. Spatial turnover of soil viral populations and genotypes
393 overlain by cohesive responses to moisture in grasslands. BioRxiv (2022)
394 doi:10.1101/2022.03.24.485562.
395 9. Albright, M. B. N. et al. Experimental evidence for the impact of soil viruses on carbon
396 cycling during surface plant litter decomposition. ISME COMMUN. 2, 24 (2022).
397 10. Durham, D. M. et al. Substantial differences in soil viral community composition within
398 and among four Northern California habitats. ISME COMMUN. 2, 100 (2022).
399 11. Braga, L. P. P. et al. Impact of phages on soil bacterial communities and nitrogen
400 availability under different assembly scenarios. Microbiome 8, 52 (2020).
401 12. Wu, R. et al. Moisture modulates soil reservoirs of active DNA and RNA viruses.
402 Commun. Biol. 4, 992 (2021).
403 13. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K.
404 Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon
405 cycling in soil. Proc Natl Acad Sci USA 116, 25900–25908 (2019).
406 14. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. Nat. Rev.
407 Microbiol. (2022) doi:10.1038/s41579-022-00811-z.
408 15. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat.
409 Microbiol. 3, 870–880 (2018).
410 16. Suttle, C. A. Marine viruses--major players in the global ecosystem. Nat. Rev. Microbiol.
411 5, 801–812 (2007).
412 17. Trubl, G. et al. Active virus-host interactions at sub-freezing temperatures in Arctic peat
413 soil. Microbiome 9, 208 (2021).

414 18. Lee, S. et al. Methane-derived carbon flows into host-virus networks at different trophic
415 levels in soil. Proc Natl Acad Sci USA 118, (2021).
416 19. Santos-Medellin, C. et al. Viromes outperform total metagenomes in revealing the
417 spatiotemporal patterns of agricultural soil viral communities. ISME J. 15, 1956–1970 (2021).
418 20. Snyder, J. C. et al. Virus movement maintains local virus population diversity. Proc Natl
419 Acad Sci USA 104, 19102–19107 (2007).
420 21. Hesse, U. et al. Virome assembly and annotation: A surprise in the namib desert. Front.
421 Microbiol. 8, 13 (2017).
422 22. Jin, M. et al. Diversities and potential biogeochemical impacts of mangrove soil viruses.
423 Microbiome 7, 58 (2019).
424 23. Trubl, G. et al. Soil viruses are underexplored players in ecosystem carbon processing.
425 mSystems 3, (2018).
426 24. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant
427 ocean viruses. Nature 537, 689–693 (2016).
428 25. Sokol, N. W. et al. Life and death in the soil microbiome: how ecological processes
429 influence biogeochemistry. Nat. Rev. Microbiol. 20, 415–430 (2022).
430 26. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the
431 marine microbial realm. Nat. Microbiol. 3, 754–766 (2018).
432 27. Liang, X. et al. Lysogenic reproductive strategies of viral communities vary with soil
433 depth and are correlated with bacterial diversity. Soil Biol. Biochem 144, 107767 (2020).
434 28. Williamson, K. E., Radosevich, M., Smith, D. W. & Wommack, K. E. Incidence of
435 lysogeny within temperate and extreme soil environments. Environ. Microbiol. 9, 2563–2574
436 (2007).
437 29. Wu, R. et al. Structural characterization of a soil viral auxiliary metabolic gene product -
438 a functional chitosanase. Nat. Commun. 13, 5485 (2022).
439 30. Wu, R. et al. DNA Viral Diversity, Abundance, and Functional Potential Vary across
440 Grassland Soils with a Range of Historical Moisture Regimes. MBio 12, e0259521 (2021).
441 31. Liang, X. et al. Viral abundance and diversity vary with depth in a southeastern United
442 States agricultural ultisol. Soil Biol. Biochem 137, 107546 (2019).
443 32. Jansson, J. K. Soil viruses: Understudied agents of soil ecology. Environ. Microbiol. 25,
444 143–146 (2023).
445 33. Hengl, T. et al. SoilGrids250m: Global gridded soil information based on machine
446 learning. PLoS ONE 12, e0169748 (2017).
447 34. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG).
448 Nat. Biotechnol. 37, 29–37 (2019).
449 35. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human
450 gut microbiome. Nat. Microbiol. 6, 960–970 (2021).
451 36. Roux, S. et al. IMG/VR v3: an integrated ecological and evolutionary framework for
452 interrogating genomes of uncultivated viruses. Nucleic Acids Res. 49, D764–D775 (2021).

453 37. Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L. & Davidson, A. R. The phage
454 lambda major tail protein structure reveals a common evolution for long-tailed phages and the
455 type VI bacterial secretion system. Proc Natl Acad Sci USA 106, 4160–4165 (2009).
456 38. Lasker, K., Mann, T. H. & Shapiro, L. An intracellular compass spatially coordinates cell
457 cycle modules in Caulobacter crescentus. Curr. Opin. Microbiol. 33, 131–139 (2016).
458 39. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z.
459 Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during
460 the infection process. Appl. Microbiol. Biotechnol. 101, 3103–3119 (2017).
461 40. Kintz, E. et al. A BTP1 prophage gene present in invasive non-typhoidal Salmonella
462 determines composition and length of the O-antigen of the lipopolysaccharide. Mol. Microbiol.
463 96, 263–275 (2015).
464 41. Carini, P. et al. Relic DNA is abundant in soil and obscures estimates of soil microbial
465 diversity. Nat. Microbiol. 2, 16242 (2016).
466 42. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial
467 diversity. Nature 551, 457–463 (2017).
468 43. Locey, K. J. et al. Dormancy dampens the microbial distance-decay relationship. Philos.
469 Trans. R. Soc. Lond. B Biol. Sci. 375, 20190243 (2020).
470 44. CAZy - GT4. http://www.cazy.org/GT4.html.
471 45. CAZy - GH73. http://www.cazy.org/GH73.html.
472 46. CAZy - CBM50. http://www.cazy.org/CBM50.html.
473 47. Wommack, K. E. & Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems.
474 Microbiol. Mol. Biol. Rev. 64, 69–114 (2000).
475 48. Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. Science 332,
476 106–109 (2011).
477 49. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and
478 evolutionary determinants of bacteriophage host range. Trends Microbiol. 27, 51–63 (2019).
479 50. Heineman, R. H., Springman, R. & Bull, J. J. Optimal foraging by bacteriophages
480 through host avoidance. Am. Nat. 171, E149-57 (2008).
481 51. Holtzman, T. et al. A continuous evolution system for contracting the host range of
482 bacteriophage T7. Sci. Rep. 10, 307 (2020).
483 52. Sant, D. G., Woods, L. C., Barr, J. J. & McDonald, M. J. Host diversity slows
484 bacteriophage adaptation by selecting generalists over specialists. Nat. Ecol. Evol. 5, 350–359
485 (2021).
486 53. Paez-Espino, D. et al. Uncovering Earth's virome. Nature 536, 425–430 (2016).
487 54. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and
488 its constraints on virus evolution. Nat. Rev. Microbiol. 17, 321–328 (2019).
489 55. Srinivasiah, S. et al. Phages across the biosphere: contrasts of viruses in soil and aquatic
490 environments. Res. Microbiol. 159, 349–357 (2008).
491 56. Srinivasiah, S. et al. Dynamics of autochthonous soil viral communities parallels
492 dynamics of host communities under nutrient stimulation. FEMS Microbiol. Ecol. 91, (2015).

493 57. Romine, M. F. et al. Elucidation of roles for vitamin B12 in regulation of folate,
494 ubiquinone, and methionine metabolism. Proc Natl Acad Sci USA 114, E1205–E1214 (2017).
495 58. Mattenberger, Y., Mattson, S., Métrailler, J., Silva, F. & Belin, D. 55.1, a gene of
496 unknown function of phage T4, impacts on Escherichia coli folate metabolism and blocks DNA
497 repair by the NER. Mol. Microbiol. 82, 1406–1421 (2011).
498 59. Sgro, G. G. et al. Bacteria-Killing Type IV Secretion Systems. Front. Microbiol. 10, 1078
499 (2019).
500 60. Cascales, E. & Christie, P. J. Definition of a bacterial type IV secretion pathway for a
501 DNA substrate. Science 304, 1170–1173 (2004).
502 61. Tan, Y. et al. Organic fertilizers shape soil microbial communities and increase soil
503 amino acid metabolites content in a blueberry orchard. Microb. Ecol. 85, 232–246 (2023).
504 62. Wang, D., Chadwick, D. R., Hill, P. W., Ge, T. & Jones, D. L. Rapid microbial uptake
505 and mineralization of 14C-labelled cysteine and methionine along a grassland productivity
506 gradient. Soil Biol. Biochem 180, 109022 (2023).
507 63. Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the
508 dynamo machine: the evolution of the F-type and V-type ATPases. Nat. Rev. Microbiol. 5, 892–
509 899 (2007).
510 64. Rao, V. B. & Feiss, M. The bacteriophage DNA packaging motor. Annu. Rev. Genet. 42,
511 647–681 (2008).
512 65. Kühlbrandt, W. Structure and Mechanisms of F-Type ATP Synthases. Annu. Rev.
513 Biochem. 88, 515–549 (2019).
514 66. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems.
515 Curr. Opin. Microbiol. 31, 161–168 (2016).
516 67. Daniel, R. The metagenomics of soil. Nat. Rev. Microbiol. 3, 470–478 (2005).
517 68. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T.
518 Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems 3,
519 (2018).
520 69. Van Twest, R. & Kropinski, A. M. Bacteriophage enrichment from water and soil.
521 Methods Mol. Biol. 501, 15–21 (2009).
522 70. Trubl, G. et al. Towards optimized viral metagenomes for double-stranded and single-
523 stranded DNA viruses from challenging soils. PeerJ 7, e7265 (2019).
524 71. Zablocki, O. et al. VirION2: a short- and long-read sequencing and informatics workflow
525 to study the genomic diversity of viruses in nature. PeerJ 9, e11088 (2021).
526 72. Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and
527 abundant viruses. Nat. Commun. 8, 15892 (2017).
528 73. Guerra, C. A. et al. Blind spots in global soil biodiversity and ecosystem function
529 research. Nat. Commun. 11, 3870 (2020).
530 74. Mukherjee, S. et al. Genomes OnLine Database (GOLD) v.8: overview and updates.
531 Nucleic Acids Res. 49, D723–D733 (2021).

532 75. Chen, I.-M. A. et al. IMG/M v.5.0: an integrated data management and comparative
533 analysis system for microbial genomes and microbiomes. Nucleic Acids Res. 47, D666–D677
534 (2019).
535 76. Clum, A. et al. DOE JGI metagenome workflow. mSystems 6, (2021).
536 77. Nelson, W. C. et al. Terabase metagenome sequencing of grassland soil microbiomes.
537 Microbiol. Resour. Announc. 9, (2020).
538 78. El-Gebali, S. et al. The Pfam protein families database in 2019. Nucleic Acids Res. 47,
539 D427–D432 (2019).
540 79. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer
541 based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5, 69
542 (2017).
543 80. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. Nat.
544 Biotechnol. (2023) doi:10.1038/s41587-023-01953-y.
545 81. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-
546 assembled viral genomes. Nat. Biotechnol. 39, 578–585 (2021).
547 82. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
548 using DIAMOND. Nat. Methods 18, 366–368 (2021).
549 83. Van Dongen, S. Graph clustering via a discrete uncoupling process. SIAM. J. Matrix
550 Anal. & Appl. 30, 121–141 (2008).
551 84. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. Nat. Biotechnol. 39, 499–
552 509 (2021).
553 85. Almeida, A. et al. A unified catalog of 204,938 reference genomes from the human gut
554 microbiome. Nat. Biotechnol. 39, 105–114 (2021).
555 86. Pasolli, E. et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over
556 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 176, 649-
557 662.e20 (2019).
558 87. Merrill, B. D. et al. Ultra-deep Sequencing of Hadza Hunter-Gatherers Recovers
559 Vanishing Gut Microbes. BioRxiv (2022) doi:10.1101/2022.03.30.486478.
560 88. Camargo, A. P. et al. IMG/VR v4: an expanded database of uncultivated virus genomes
561 within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids
562 Res. 51, D733–D743 (2023).
563 89. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic
564 niche specialization in the "core" and "flexible" Pacific Ocean Virome. ISME J. 9, 472–484
565 (2015).
566 90. R Core Team. R: A language and environment for statistical computing. (R Foundation
567 for Statistical Computing, 2021).
568 91. Wickham, H. & Chang, W. Package "ggplot2." Create elegant data … (2016).
569 92. Wickham, H. reshape2: flexibly reshape data: a reboot of the reshape package.
570 93. Kolde, R. & Kolde, M. R. Package "pheatmap." R Package (2018).
571 94. Jr, F. H. & Jr, M. H. Package "hmisc." CRAN2018 (2019).

95. Goslee, S. C. & Urban, D. L. The ecodist Package for Dissimilarity-based Analysis of Ecological Data. J. Stat. Softw. 22, (2007).
96. Neuwirth, E. & Brewer, R. C. ColorBrewer palettes. R package version (2014).
97. Becker, R. A., Wilks, A. R., Brownrigg, R. & Minka, T. P. maps: Draw geographical maps.
98. R: The R Stats Package. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html.
99. Hijmans, R. J., Williams, E., Vennes, C. & Hijmans, M. R. J. Package "geosphere."
100. Wickham, H. Tools for Splitting, Applying and Combining Data [R package plyr version 1.8.8]. (2022).
101. A Grammar of Data Manipulation [R package dplyr version 1.1.2]. https://cran.r-project.org/web/packages/dplyr/index.html (2023).
102. Wickham, H. Simple, Consistent Wrappers for Common String Operations [R package stringr version 1.5.0]. (2022).
103. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE 8, e61217 (2013).
104. phyloT : a phylogenetic tree generator. https://phylot.biobyte.de/.
105. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36 (2016).
106. Wang, L.-G. et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol. Biol. Evol. 37, 599–603 (2020).
107. Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in'ggplot2'. R package version 0.4 (2020).
108. DE Beaudette, Skovlin, J., Roecker, S. & Beaudette, M. D. E. Package "soilDB." (2022).
109. Luo, W. & Brouwer, C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29, 1830–1831 (2013).

599 **ONLINE METHODS**

600 *Data Collection and Curation.*

601 We collected a total of 2,953 soil metagenomic samples from soils from major
602 repositories and ecological networks including the JGI Integrated Microbial Genomes and
603 Microbiomes (IMG/M) platform, MG-RAST metagenomics analysis server, Global Urban Soil
604 Ecological Education Network (GLUSEEN), Earth Microbiome Project (EMP), and National
605 Ecological Observatory Network (NEON) plus submissions from individual collaborators. This
606 included 1,552 samples not previously included in IMG/M (Fig. 1 and 2). All dataset authors
607 were contacted for data reuse permissions.

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

622 were manually excluded. This resulted in 1,480 curated metagenomes from publicly available
623 data in IMG/M.
624 After collating samples from JGI IMG/M and the newly collected samples from external
625 networks and collaborators, the final dataset consisted of 2,953 soils with 2,015,688,128 contigs,
626 representing 1.2 terabases of assembled DNA sequences.
627 In parallel, we retrieved mean values for soil parameters from the SoilGrids250m
628 database from 0-5 cm[33]. SoilGrids250m is a spatial interpolation of global soil parameters using
629 ~150,000 soil samples and 158 remote sensing-based products. Here, we focus on six
630 parameters often associated with soil microbial communities: bulk density, cation exchange
631 capacity, nitrogen, pH, soil organic carbon, and clay content. Because our focus on spatial
632 dynamics and soils were collected at various times, we did not include temporally dynamic
633 variables such as soil moisture or temperature in our set of environmental parameters, though we
634 acknowledge they may have profound impacts on the soil virosphere.
635
636 *Assembly and Annotation of Samples added to IMG/M.*
637 To standardize data analysis across all samples, the 1,552 soil metagenomic samples not
638 collected from IMG/M were analyzed using the JGI's Metagenome Workflow [75]. In brief,
639 samples were individually assembled using MetaSpades v3.1. 1,476 of the 1,552 assembled soil
640 samples passed default quality control thresholds[76], yielding 133 gigabases of assembled DNA in
641 241,465,924 contigs. Additionally, three very large metagenomes (>1TB each) were assembled
642 separately due to computational limitations in standard workflows[77]. The resulting assemblies
643 were assigned GOLD identification numbers and imported into IMG/M and processed using

644 IMG/M Metagenome Annotation Pipeline v5.0.0 to align with data obtained directly from
645 IMG/M[75].
646
647 *Virus identification, clustering, and host prediction.*
648 We performed an initial identification of viral contigs using using a modified version of
649 the IMG/VR v3's virus identification pipeline (code available at
650 https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline)[35,36]. The pipeline
651 identifies viruses based on the presence of 23,841 virus protein families, 16,260 protein families
652 of microbial origin from the Pfam database[78] and VirFinder[79] to identify putative viral genomes
653 in contigs that were at least 1 kb long. During the course of this study, geNomad v1.3.3[80], a tool
654 for virus identification with improved classification performance was released and incorporated
655 into our pipeline to improve prediction confidence and perform taxonomic assignment. We
656 further processed predicted viral sequences using CheckVv1.0.1 (database version 1.5)[81] to
657 assess the quality of the viral genomes. As this study focused on non-integrated virus genomes,
658 contigs that were flagged by either geNomad or CheckV as proviral were discarded. From the
659 remaining contigs, virus genomes were selected using the following rules: (1) contigs of at least
660 1kb with high similarity to genomes in the CheckV database (that is, that had high- or medium-
661 quality completeness estimates) or that contained direct terminal repeats were automatically
662 selected; (2) contigs longer than 10kb were required to have a geNomad virus score higher than
663 0.8 and to either encode one virus hallmark (e.g., terminase, capsid proteins, portal protein, etc.),
664 as determined by geNomad, or to have a geNomad virus marker of at least 5.0; (3) contigs
665 shorter than 10 kb and longer than 5 kb were required to have a geNomad virus score higher than
666 0.9, to encode at least one virus hallmark, and to have a virus marker enrichment higher than 2.0.

667 This resulted in 49,649 viral contigs that we used for downstream analysis. All viral contigs are
668 available at: https://doi.org/10.25584/2229733.
669 Viral genomes were clustered into viral operational taxonomic units (vOTUs) following
670 MIUViG guidelines (95% average nucleotide identity, ANI; 85% aligned fraction[34]). In brief, we
671 performed an all-versus-all BLAST (v2.13.0+, '-task megablast -evalue 1e-5 -max_target_seqs
672 20000') search to estimate pairwise ANIs and AFs, as described in Nayfach et al.[81] and
673 employed pyLeiden (available at: https://github.com/apcamargo/pyleiden) to cluster genomes,
674 using as input a graph where pairs of genomes that satisfied the MIUViG criteria were connected
675 by edges. Viruses were also grouped at approximate genus-level (40% AAI — average amino
676 acid identity; 20% shared genes) and family-level (20% AAI; 10% shared genes) clusters using
677 DIAMOND[82] for protein alignment and Markov Cluster Process (MCL)[83] for clustering[35].
678 Viral sequences were assigned to putative host (bacterial and archaeal) taxa through
679 matches to a previously described database of CRISPR spacers of 1.6 million bacterial and
680 archaeal genomes from NCBI GenBank and MAGs (release 242; 15 February 2021)[84–88].
681 Sequences of viral genomes were queried against the spacer database
682 (https://portal.nersc.gov/cfs/m342/crisprDB) using blastn (v2.9.0+, parameters: '-
683 max_target_seqs = 1000 -word_size = 8 -dust = no'). Only alignments with at least 25 bp, less
684 than 2 mismatches, and that covered ≥95% of the spacer length were considered. Viral sequences
685 were assigned to the host taxon at the lowest taxonomic rank that had at least two spacers
686 matched and that represented >70% of all matches.
687
688 *Potential Auxiliary Metabolic Gene (AMG) Prediction.*

689 We leveraged an intermediate output of geNomad (v1.3.3)[80] ('genes.tsv') to screen
690 putative AMGs on the detected viral contigs. Proteins of the viral contigs were annotated by
691 virus and host-specific markers implemented in geNomad. The identified viral hallmark (e.g.,
692 terminase, major capsid protein) and non-hallmark proteins were labeled as 'VV-1' and 'V*-0'
693 in geNomad output, respectively. The rest of the viral proteins of the detected viral contigs that
694 were annotated as non-virus-specific or unclassified were then classified into five categories of
695 putative AMGs based on the presence of viral hallmark or non-hallmarks up- or down-stream as
696 mentioned previously[30]. The AMGs with both virus-specific genes ('VV-1' or 'V*-0') were
697 retained for the following analysis. To improve the functional annotations of the putative AMGs
698 and highlight the viral potentials of metabolizing carbohydrates and glycoconjugates, the AMG
699 proteins were also annotated by Carbohydrate-Active enZYmes (CAZY) Database and Kyoto
700 Encyclopedia of Genes and Genomes (KEGG) database using the default settings in addition to
701 the functional annotation databases implemented in geNomad. The putative AMG was assigned
702 to the functional annotation with the highest bitscore (e.g,. duplicate annotations were not
703 allowed). Following Hurwitz and U'Ren[66] and Hurwitz et al[89], we further screened putative
704 AMGs to remove genes not found in KEGG pathways. Additionally, in recognition of the
705 ambiguity in distinguishing genes encoding auxiliary metabolic functions versus core metabolic
706 processes[66], we discuss the resulting set of genes presented here as 'putative AMGs'.
707
708 *Statistical Analysis.*
709
710 All statistical analyses and data visualizations were performed using *R* v4.1.0[90]. We used
711 the following packages for data manipulation and visualization: ggplot2[91], reshape2[92],

712 pheatmap[93], Hmisc[94], ggpubr[95], RColorBrewer[96], maps[97], stats[98] geosphere[99], plyr[100], dplyr[101],
713 and stringr[102]. Packages pertaining to specific analyses are listed below.

714 We generated rarefaction curves for individual samples and for cumulative sequencing
715 depth (Fig. S1 and S2) using the 'phyloseq' package [103] and custom R plots, respectively.
716 Samples containing less than five vOTUs, viral clusters, or viral Pfams; or less than 100
717 CRISPR-spacer-based host taxon, microbial Pfams, or microbial taxa were removed for visual
718 clarity. Removed samples followed the same general trends as presented in Fig. S1. To visualize
719 saturation across cumulative sequencing depth (Fig. S2), we ordered samples from lowest to
720 highest total assembled bp and progressively added them along the x-axis. On the y-axis, we plot
721 the associated cumulative number of unique attributes.

722 A phylogenetic tree of CRISPR-spaced-based host taxon was generated at the order-level
723 using phyloT v2[104], an online tree generator based on the Genome Taxonomy Database. Then,
724 we visualized the tree in R using the packages 'ggtree'[105], 'treeio'[106], and 'ggnewscale'[107]. To
725 examine relationships between common microbial hosts of soil viruses and soil properties, we
726 first downloaded data describing bulk density, cation exchange capacity, nitrogen, pH, soil
727 organic carbon, and clay content from the SoilGrids250m database[33] using the 'soilDB'
728 package[108]. Mean values of soil properties from 0-5 cm were correlated to the total number of
729 CRISPR spacer hits per microbial ordered using Spearman correlation.

730 Finally, we mapped genes detected across the entire soil virosphere (i.e., all samples
731 combined) to their corresponding KEGG pathways using the 'pathview' package in R[109]. Gene
732 abundances were converted by log base 10 for visualization.

733

734 **DATA AVAILABILITY.**

735 The GSV Atlas is available for download at: https://doi.org/10.25584/2229733. It includes all

736 UViGs regardless of quality (File 1, 616,935 UViGs), a file containing data associated with each
737 contig that passed QA/QC (File 2, 49,649 contigs), predicted viral protein sequences (File 3,
738 402,882 predicted protein sequences), a file containing data associated with each gene (File 4,
739 1,432,147 genes), geographic and physicochemical data of the curated soil samples (File 5, 2,953
740 samples), and a readme file (File 6).
741
742 **CODE AVAILABILITY.** Code for sequence processing is described in Nayafch et al.[35] and is
743 available in the materials associated with those publications. Github repositories associated with
744 this publication are available at:
745 https://github.com/snayfach/MGV/tree/master/viral_detection_pipeline and
746 https://github.com/apcamargo/pyleiden). All other code is available upon request.
747
748

CONSORTIUM AUTHORS.

Name; Affiliation; E-mail; ORCID

Emily B. Graham; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA; School of Biological Sciences, Washington State University, Pullman, WA emily.graham@pnnl.gov; 0000-0002-4623-7076

Antonio Pedro Camargo; US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; antoniop.camargo@lbl.gov; 0000-0003-3913-2484

Ruonan Wu; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA; ruonan.wu@pnnl.gov; 0000-0001-9466-4462

Russell Y. Neches; US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ryneches@gmail.com; 0000-0002-2055-8381

Matt Nolan; US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; mpnolan@lbl.gov; 0000-0002-9532-0853

David Paez-Espino; US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; paezespino@gmail.com; 0000-0002-2939-5398

Nikos C. Kyrpides; US Department of Energy Joint Genome Institute, Lawrence Berkeley

National Laboratory, Berkeley, CA, USA; nckyrpides@lbl.gov; 0000-0002-6131-0462

Janet K. Jansson; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA; janet.rae.jansson@gmail.com; 0000-0002-5487-4315

Jason E. McDermott; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA; jason.mcdermott@pnnl.gov; 0000-0003-2961-2572

Kirsten S. Hofmockel; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA USA; kirsten.hofmockel@pnnl.gov; 0000-0003-1586-2167

National Ecological Observatory Network; NEON, Boulder, CO USA; none; none

Jeffrey L Blanchard; Biology, University of Massachusetts Amherst, Leverett, MA USA; jlb@umass.edu; 0000-0002-7310-9678

Xiao Jun A. Liu; Institute for Environmental Genomics, University of Oklahoma, Norman, OK, USA; xj.allen.liu@gmail.com; 0000-0002-0488-8591

Jorge L. Mazza Rodrigues; Department of Land, Air and Water Resources, University of California - Davis, Davis, CA, USA; jmrodrigues@ucdavis.edu; 0000-0002-6446-6462

Zachary B. Freedman; Department of Soil Science, University of Wisconsin-Madison, Madison, WI, USA; zfreedman@wisc.edu; 0000-0001-9160-7470

Petr Baldrian; Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic; baldrian@biomed.cas.cz; 0000-0002-8983-2721

Martina Stursova; Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic; stursova@biomed.cas.cz; 0000-0003-1387-6426

Kristen M. DeAngelis; Microbiology Department, University of Massachusetts Amherst, Amherst, MA, USA; deangelis@microbio.umass.edu; 0000-0002-5585-4551

Sungeun Lee; Ecole Centrale de Lyon, Ecully, France; sungeun.lee@ec-lyon.fr; 0000-0001-8878-6463

Filipa Godoy-Vitorino; Department of Microbiology and Medical Zoology, University of Puerto Rico School of Medicine, Medical Sciences Campus, PO Box 365067, 00936 San Juan, Puerto Rico, USA; filipa.godoy@upr.edu; 0000-0003-1880-0498

Yun Kit Yeoh; Australian Institute of Marine Science, Townsville, Queensland, Australia; yun.yeoh@uqconnect.edu.au; 0000-0003-0241-6117

Hinsby Cadillo-Quiroz; School of Life Sciences, Arizona State University; hinsby@asu.edu;

818 0000-0002-4908-4597

819

820 Susannah G. Tringe; Environmental Genomics and Systems Biology, Lawrence Berkeley

821 National Laboratory, Berkeley, CA, USA; SGTringe@lbl.gov; 0000-0001-6479-8427

822

823 Archana Chauhan; Panjab University, Chandigarh, India; achauhan@pu.ac.in;

824 achauhan@puchd.ac.in; 0000-0002-6821-0464

825

826 Don A Cowan; Centre for Microbial Ecology and Genomics, Department of Biochemistry,

827 Genetics and Microbiology, University of Pretoria, Pretoria, South Africa; don.cowan@up.ac.za;

828 0000-0001-8059-861X

829

830 Marc W. Van Goethem; Centre for Microbial Ecology and Genomics, Department of

831 Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa;

832 mwvangoethem@gmail.com; 0000-0001-8688-3323

833

834 Tanja Woyke; U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National

835 Laboratory, Berkeley, California, USA; twoyke@lbl.gov; 0000-0002-9485-5637

836

837 Nicholas C. Dove; University of California, Merced--Merced, CA USA; ndove7@gmail.com;

838 0000-0003-1152-956x

839

840 Konstantinos T. Konstantinidis; School of Civil and Environmental Engineering, and School of

Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA.; kostas@ce.gatech.edu; 0000-0002-0954-4755

Thomas E. Juenger; Department of Integrative Biology, University of Texas, Austin, Texas; tjuenger@austin.utexas.edu; 0000-0001-9550-9288

Stephen C. Hart; Department of Life and Environmental Sciences and the Sierra Nevada Research Institute, University of California, Merced, CA, USA; shart4@ucmerced.edu; 0000-0002-9023-6943

David D. Myrold; Department of Crop and Soil Science, Oregon State University, Corvallis, OR, USA; n/a; n/a

Tullis C. Onstott; Department of Geosciences, Princeton University, Princeton, NJ, USA; n/a; n/a

Brendan J.M. Bohannan; Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA; bohannan@uoregon.edu; 0000-0003-2907-1016

Marty R. Schmer; United States Department of Agriculture, Agricultural Research Service, Lincoln, NE; marty.schmer@usda.gov; 0000-0002-3721-6177

Nathan A. Palmer; Wheat, Sorghum, and Forage Research Unit. United States Department of

Agriculture - Agricultural Research Service. Lincoln, NE. USA; nathan.palmer@usda.gov; 0000-0001-8088-6371

Klaus Nusslein; Department of Microbiology, University of Massachusetts - Amherst, MA; nusslein@microbio.umass.edu; 0000-0002-0663-4448

Thulani P. Makhalanyane; DSI/NRF SARChI in Marine Microbiomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria; thulani.makhalanyane@up.ac.za; 0000-0002-8173-1678

Katherine A. Dynarski; Department of Land, Air, and Water Resources, University of California Davis, Davis, California, USA; katherine.dynarski@usda.gov; 0000-0001-5101-9666

Neslihan Tas; Climate and Ecosystem Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA; ntas@lbl.gov; 0000-0001-7525-2331

Graeme W. Nicol; Laboratoire Ampère, Ecole Centrale de Lyon, Ecully, France; graeme.nicol@cnrs.fr; 0000-0002-3876-022X

Christina Hazard; Laboratoire Ampère, Ecole Centrale de Lyon, Ecully, France; christina.hazard@ec-lyon.fr; 0000-0002-0325-5856

Erin D. Scully; USDA-ARS Center for Grain and Animal Health Research Manhattan, KS USA;

887 erin.scully@usda.gov; 0000-0002-8315-5619

888

889 Kunal R. Jain; Environmental Genomics and Proteomics Lab, Department of Biosciences,
890 Satellite Campus, Vadtal Raod, Sardar Patel University, Bakrol - 388 315 (Anand), Gujarat,
891 India; qnal.jain@gmail.com ; 0009-0001-6346-9055

892

893 Datta Madamwar; P. D. Patel Institute of Applied Sciences, Charotar University of Science and
894 Technology, Changa - 388 421 (Anand), Gujarat, India; datta_madamwar@yahoo.com; 0000-
895 0003-3301-1120

896

897 Andrew Bissett; Commonwealth Scientific and Industrial Research Organisation , Hobart,
898 Tasmania, Australia, 7000; Andrew.Bissett@csiro.au; 0000-0001-7396-1484

899

900 Philippe Constant; Centre Armand-Frappier Santè Biotechnologie, Institut national de la
901 recherche scientifique, Laval, Quèbec, Canada; philippe.constant@inrs.ca; 0000-0003-2739-
902 2801

903

904 Rafael S. Oliveira; Department of Plant Biology, University of Campinas, Campinas, São Paulo,
905 Brazil; rafaelsoliv@gmail.com; 0000-0002-6392-2526

906

907 Cristina Takacs-Vesbach; Department of Biology, University of New Mexico, Albuquerque,
908 New Mexico; cvesbach@unm.edu; 0000-0002-5535-2201

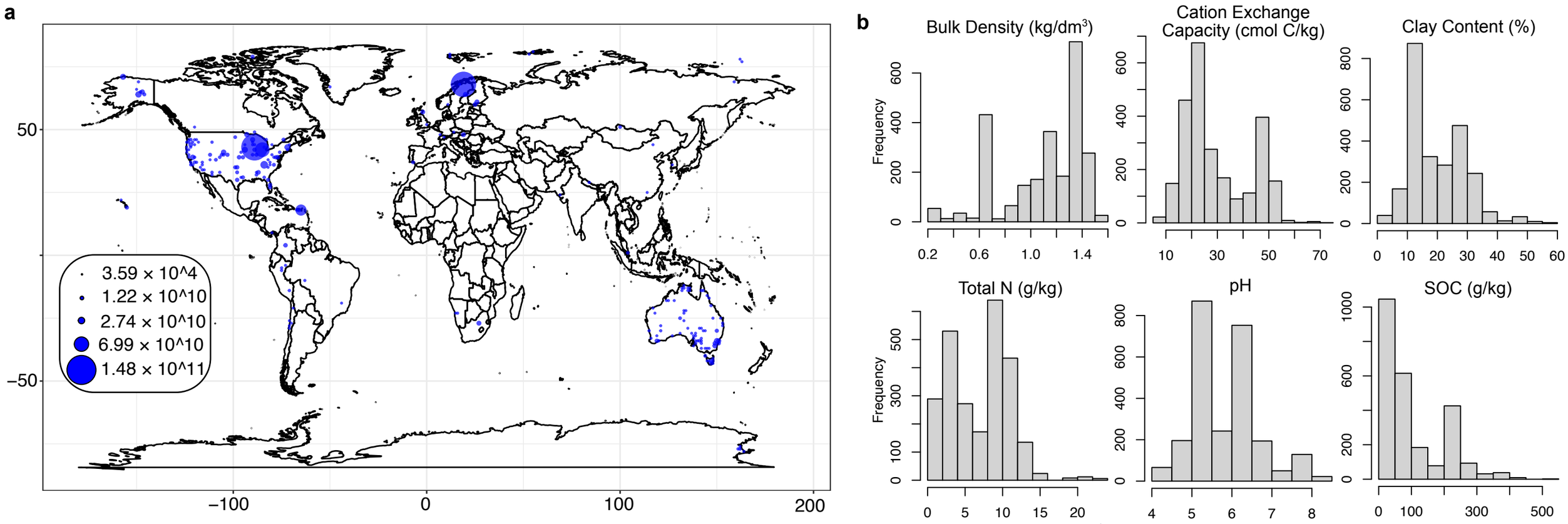
909

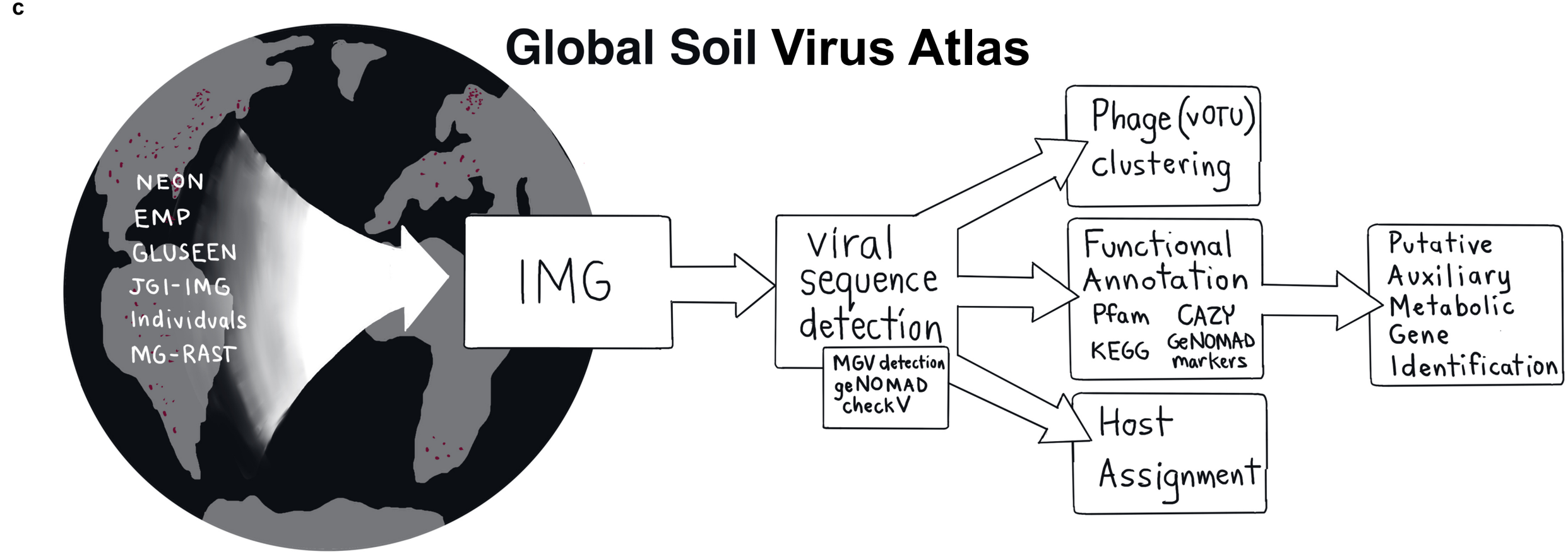
Melissa A. Cregger; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge TN; creggerma@ornl.gov; 0000-0001-8329-366X

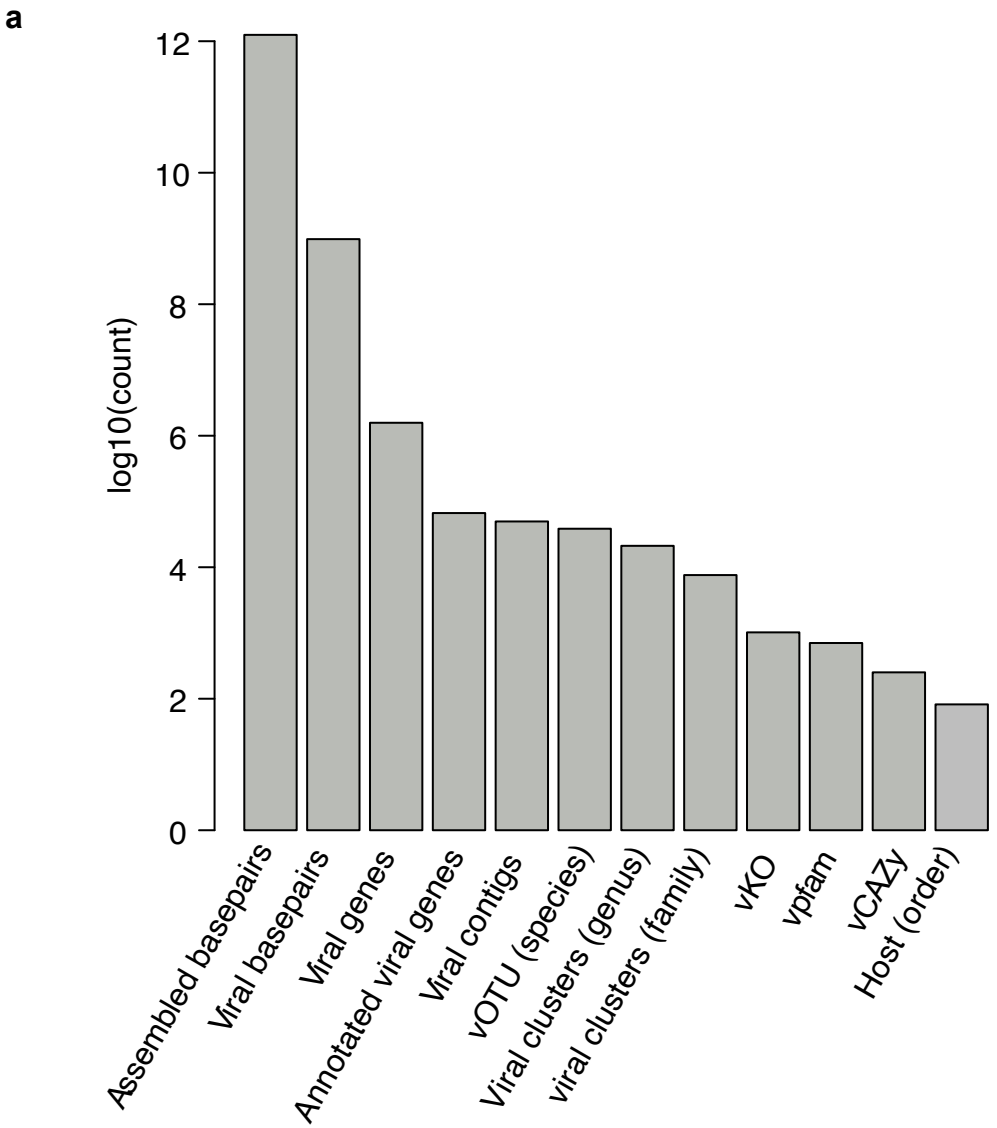
Alyssa A. Carrell; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA; carrellaa@ornl.gov; 0000-0003-1142-4709

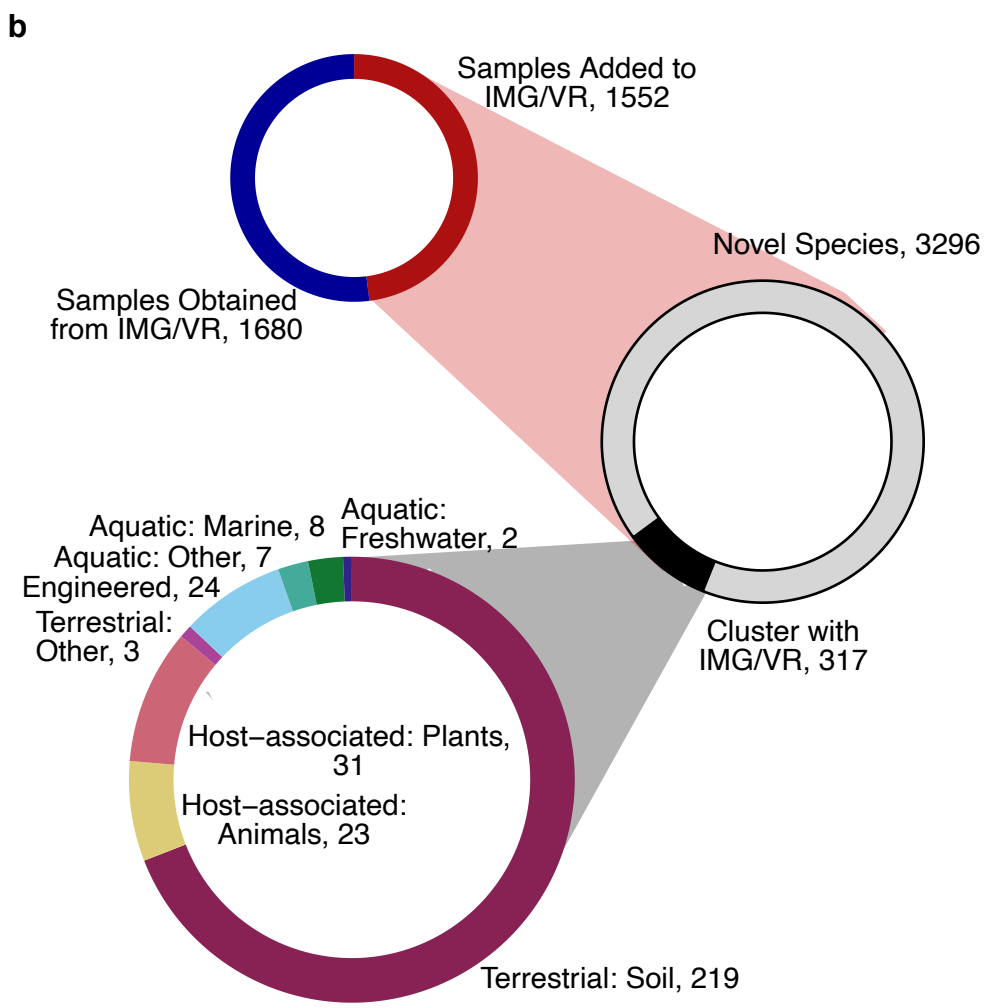
Dawn M. Klingeman; Oak Ridge National Laboratory. Oak Ridge TN 37831; klingemandm@ornl.gov; 0000-0002-4307-2560

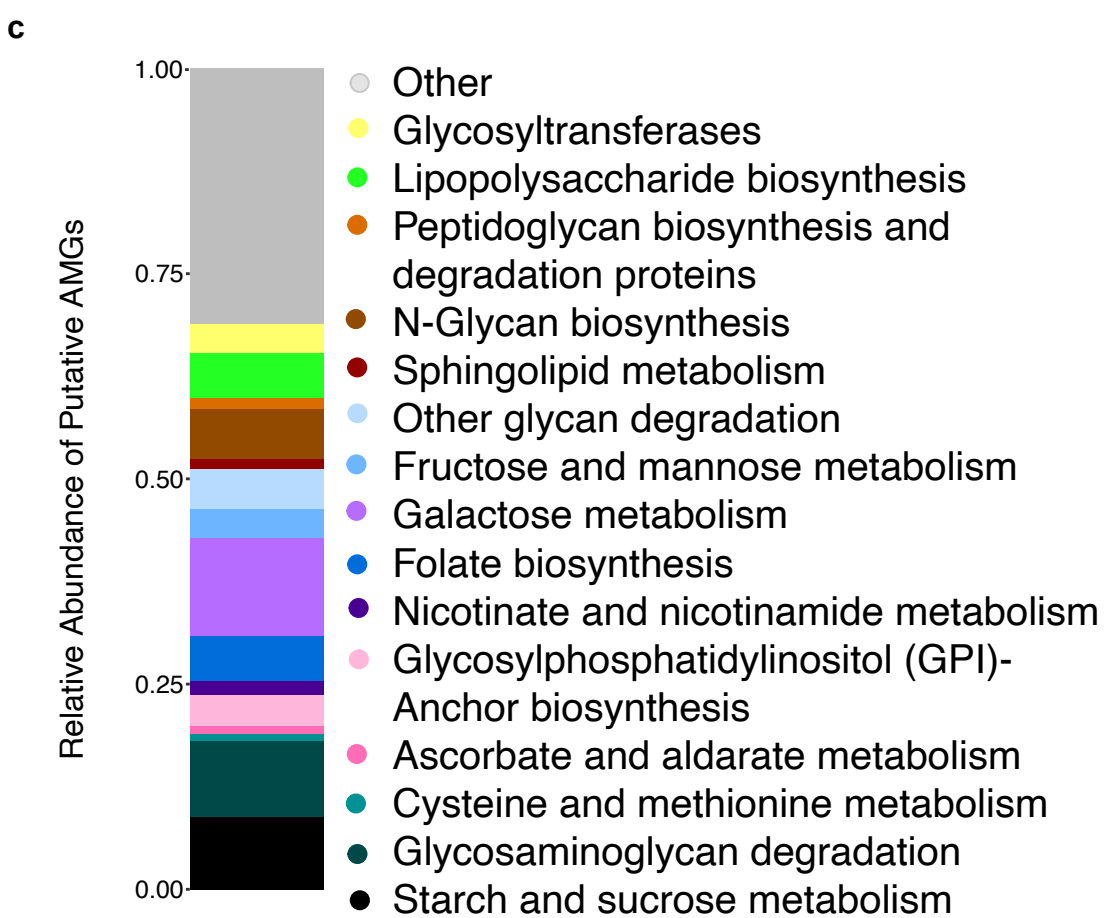
Nicole Pietrasiak; School of Life Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA; nicole.pietrasiak@unlv.edu; 0000-0003-4636-8006

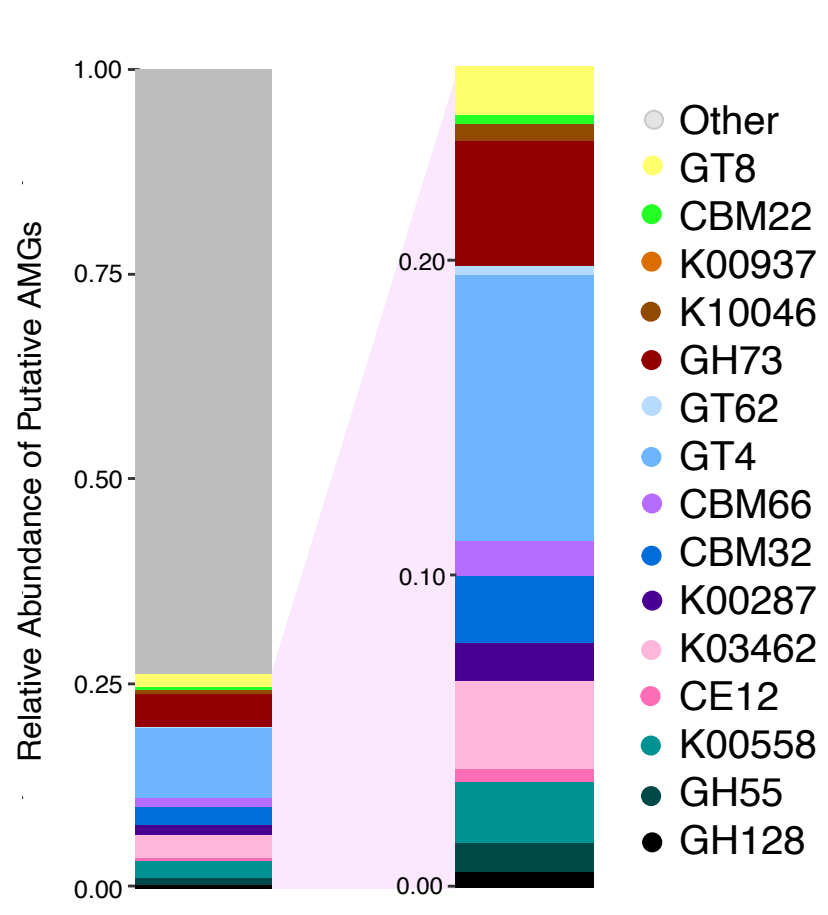

a
50
−50
−100
0
100
200
3.59 × 10^4
1.22 × 10^10
2.74 × 10^10
6.99 × 10^10
1.48 × 10^11
b
Bulk Density (kg/dm³)
Cation Exchange Capacity (cmol C/kg)
Clay Content (%)
Total N (g/kg)
pH
SOC (g/kg)
Frequency

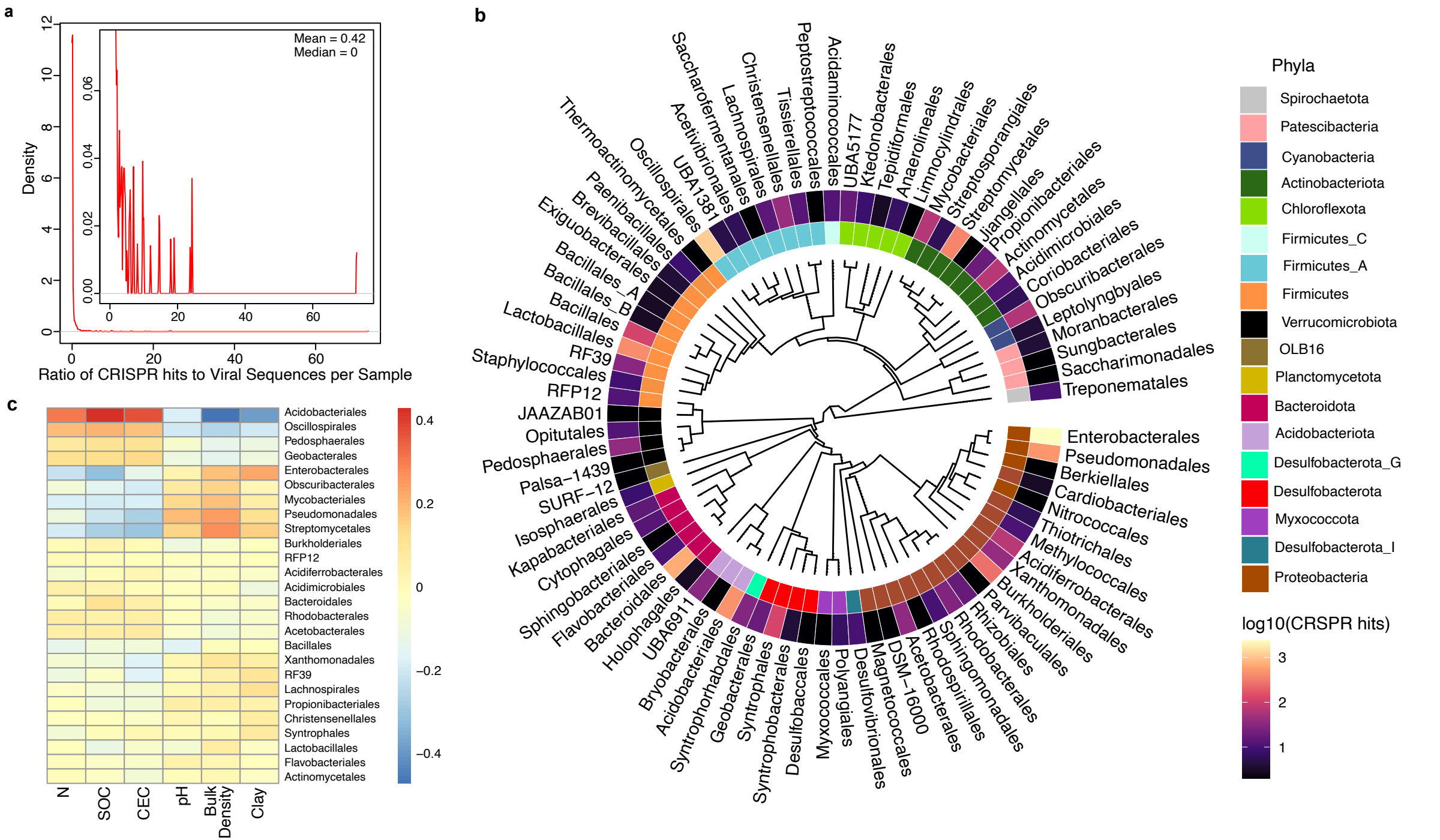

c
Global Soil Virus Atlas
NEON
EMP
GLUSEEN
JGI-IMG
Individuals
MG-RAST
IMG
Viral sequence detection
MGV detection
geNOMAD
checkV
Phage (vOTU) clustering
Functional Annotation
Pfam
CAZY
KEGG
GeNOMAD markers
Host Assignment
Putative Auxiliary Metabolic Gene Identification

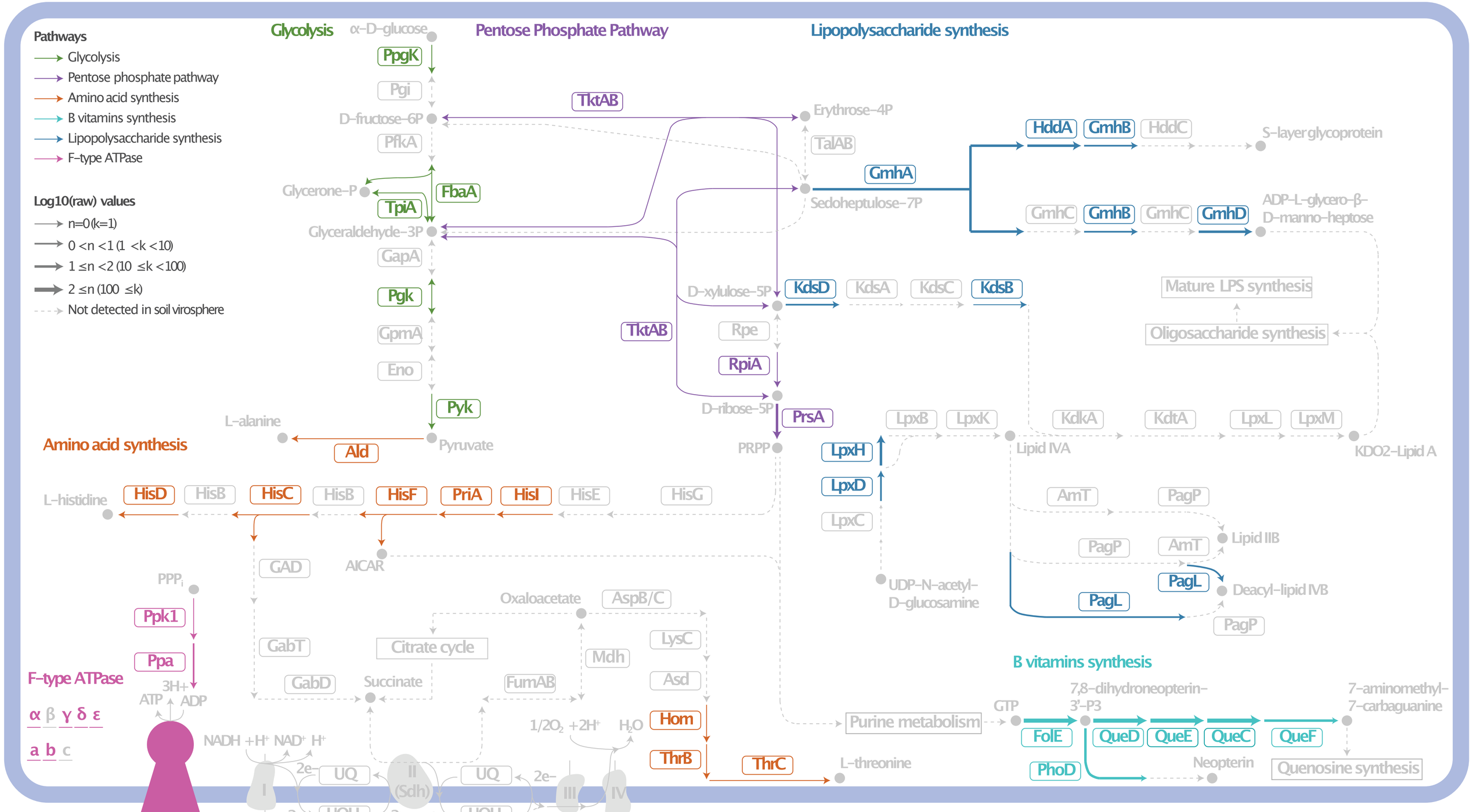

a
log10(count)
0
2
4
6
8
10
12
Assembled basepairs
Viral basepairs
Viral genes
Annotated viral genes
Viral contigs
vOTU (species)
Viral clusters (genus)
viral clusters (family)
vKO
vpfam
vCAZy
Host (order)


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+