



HAL
open science

Methane-derived carbon flows into host–virus networks at different trophic levels in soil

Sungeun Lee, Ella Sieradzki, Alexa Nicolas, Robin Walker, Mary Firestone,
Christina Hazard, Graeme Nicol

► **To cite this version:**

Sungeun Lee, Ella Sieradzki, Alexa Nicolas, Robin Walker, Mary Firestone, et al.. Methane-derived carbon flows into host–virus networks at different trophic levels in soil. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118 (32), pp.e2105124118. 10.1073/pnas.2105124118 . hal-03411995

HAL Id: hal-03411995

<https://hal.science/hal-03411995>

Submitted on 12 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Methane-derived carbon flow through host-virus trophic networks in soil

2

Sungeun Lee¹, Ella T. Sieradzki², Alexa M. Nicolas³, Robin L. Walker⁴, Mary K.
4 Firestone^{2,5}, Christina Hazard^{1*} and Graeme W. Nicol^{1*#}

6 ¹ Environmental Microbial Genomics, Laboratoire Ampère, École Centrale de Lyon,
CNRS UMR 5005, Université de Lyon, Ecully 69134, France

8 ² Department of Environmental Science, Policy and Management, University of
California, Berkeley, Berkeley, CA 94720, USA

10 ³ Department of Plant & Microbial Biology, University of California, Berkeley, Berkeley,
CA 94720, USA

12 ⁴ Scotland's Rural College, Craibstone Estate, Aberdeen, AB21 9YA, United Kingdom

⁵ Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA
14 94720, USA

*Joint senior author, # Corresponding author

16

Abstract

18 The concentration of atmospheric methane continues to increase with microbial
communities controlling soil-atmosphere fluxes. While there is substantial knowledge
20 of the diversity and function of organisms regulating methane production and
consumption, the frequency and impact of interactions with viruses on their activity in
22 soil is unknown. Metagenomic sequencing of soil microbial communities has enabled
identification of linkages between viruses and hosts. However, determining host-virus
24 linkages through sequencing does not determine whether a virus or a host are active.
In this study, we identified active individual interactions *in situ* by following the transfer
26 of assimilated carbon from active hosts to viruses. Using DNA stable-isotope probing
combined with metagenomic analyses, we characterized methane-fueled microbial
28 networks in acidic and neutral pH soils, specifically primary and secondary utilisers of
carbon, together with the recent transfer of methane-derived carbon to viruses. Sixty-
30 three percent of viral contigs from replicated soil incubations contained genes
associated with known methanotrophic bacteria. Genomic sequences from ¹³C-
32 enriched viruses were present in clustered regularly interspaced short palindromic
repeats (CRISPR) arrays of multiple, closely-related *Methylocystis* populations,

34 revealing differences in their history of viral interaction. Viruses infecting non-
methanotrophic methylotrophs and heterotrophic predatory bacteria were also
36 identified through the analysis of shared homologous genes, demonstrating that
carbon is transferred to a diverse range of viruses associated with methane-fueled
38 microbial food networks.

40 **Main text**

Microorganisms play a central role in global carbon (C) biogeochemical cycling in soil
42 systems. Soil is one of the most diverse habitats in the biosphere and can typically
contain 10^9 - 10^{10} prokaryotic cells (Frossard et al., 2016) or viruses (Williamson et al.,
44 2017) per g. Infection by viruses facilitates the horizontal transfer of genes and viral
lysis acts as a control of host abundance and releases nutrients. In the marine
46 environment, 20-40% of prokaryotes are lysed on a daily basis with the release of 150
Gt of carbon annually (Suttle, 2007). However, the role of viruses in influencing
48 prokaryotic ecology in soil remains comparatively unknown (Emerson, 2019). In
particular, difficulties remain in identifying the frequency of active interactions between
50 native host and virus populations *in situ*, largely due to a lack of tools to study
interactions within the highly complex and heterogeneous soil environment. While red-
52 queen or 'arms race' dynamics have not yet been observed in natural soil populations
as they have in marine systems (Ignacio-Espinoza et al., 2020), studies have shown
54 viruses can coevolve with their hosts in soil and that hosts change in their susceptibility
to infection (Gómez and Buckling, 2011). Shotgun sequencing of diverse soil microbial
56 communities has enabled identification of linkages between viruses and hosts
involved in carbon cycling both through identifying CRISPR spacer sequences in viral
58 genomes and the presence of viral genes encoding enzymes involved in complex
carbon degradation (Emerson et al., 2018). However, determining virus-host
60 associations *in situ* with these methods does not elucidate the frequency of viral
infections, with linkages potentially associated with populations not active under
62 current conditions, or even relic DNA (Carini et al., 2017).

Methanotrophs are a critically important group in soil systems, removing 5% of
64 atmospheric methane (Curry, 2007) and controlling fluxes to the atmosphere from
methanogenic activity in anoxic compartments (Le Mer and Roger, 2001). Aerobic
66 methanotrophs use CH_4 for both carbon and energy requirements and key
representatives in soil belong to the type I Gammaproteobacteria family

68 *Methylococcaceae*, type II Alphaproteobacteria families *Methylocystaceae* and
70 *Beijerinckiaceae*, and *Methylacidiphilaceae* of the Verrucomicrobia (Kneif, 2015). Soil
pH is one of many factors influencing methanotroph activity, with type I and type II
methanotrophs often dominating activity in neutral and acidic pH soils, respectively
72 (Zhao et al., 2020). In addition, a wide variety of non-methanotrophic methylotrophs
utilise methanol produced and excreted by methanotrophs, and together
74 methanotrophic and other methylotrophic single carbon compound (C1)-utilising
consortia assimilate methane-derived carbon in a variety of habitats (Chistoserdova
76 et al., 2010).

A widely used technique for identifying active populations within a diverse
78 microbial community in environmental samples, including methanotrophs, is DNA
stable isotope probing (SIP) (Radajewski et al., 2002). Incorporation of a substrate
80 with an enriched isotope can be traced in genomes of community members,
demonstrating utilisation of a specific substrate linked to the associated functional
82 process. As viruses are entirely composed of elements derived from a host cell, their
production inside active hosts incorporating an isotopically-enriched substrate will also
84 result in detectable viral isotopic enrichment (Pasulka et al., 2018). In this study we
aimed to identify active virus-host interactions within a complex soil habitat by
86 focussing on a taxonomically and functionally restricted group of organisms. By
following ^{13}C flow *in situ*, we aimed specifically to identify lytic DNA viruses of
88 methanotrophs actively using CH_4 -derived C, including the identification of individual
virus-host interactions, and potentially those actively infecting secondary utilisers such
90 as non-methanotrophic methylotrophs.

After aerobically incubating pH 4.5 and 7.5 soils in the presence of ^{12}C - or ^{13}C -
92 CH_4 , high buoyant density genomic DNA ($>1.732 \text{ g ml}^{-1}$) containing ^{13}C -enriched or
 ^{12}C -high GC mol% genomic DNA was recovered from triplicate incubations per isotope
94 and soil via isopycnic centrifugation in CsCl gradients (Supplementary Fig. 1). Six
metagenomes were produced from ^{13}C isotopically-enriched DNA samples only (three
96 pH 4.5, three pH 7.5; Supplementary Table 1). Concentrations of high buoyant density
genomic DNA from ^{12}C - CH_4 incubations were too low for comparable shotgun
98 sequencing. While this indicated minimal recovery of unenriched DNA in ^{13}C -
incubated samples, analysis of 16S rRNA gene amplicon libraries prepared from high
100 buoyant density DNA of both ^{12}C and ^{13}C - CH_4 incubations confirmed ^{13}C -enrichment
of C1-utilising populations (Supplementary Text, Supplementary Fig. 2).

102 Reads from individual metagenomes were assembled before taxonomic
assignment of individual contigs. Reproducibly distinct communities were enriched in
104 the two soils (Supplementary Fig. 2), with six bacterial families representing annotated
contigs >5 kbp to which >1% of reads were mapped and all including known C1-
106 utilising taxa (*Beijerinckiaceae*, *Bradyrhizobiaceae*, *Hyphomicrobiaceae*,
Methylococcaceae, *Methylocystaceae* and *Methylophilaceae*). We resolved twenty-
108 three medium and high-quality (Bowers et al., 2017) metagenome-assembled
genomes (MAGs) (Supplementary Table 2), including 12 methanotrophs. Specifically,
110 3 MAGs represented gammaproteobacterial type I methanotrophs (*Methylobacter*)
and 9 MAGs represented alphaproteobacterial type II methanotrophs (*Methylocystis*,
112 *Methylosinus* or *Methylocapsa*). Secondary utilisers of methane-derived organic
carbon were also identified with 9 MAGs associated with established or putative non-
114 methanotrophic methylotrophs, lacking methane oxidation machinery but capable of
utilising methanotroph-derived methanol (see Supplementary Text). These included
116 representatives of the *Gemmatimonadales*, *Hyphomicrobium*, *Herminiimonas* and
Rudaea, the latter two, to our knowledge, not having been previously associated with
118 methylotrophy but possessed predicted methanol and formate dehydrogenases
(Supplementary Table 2). Two MAGs represented strains of *Bdellovibrio* and
120 *Myxococcus*, known predatory bacteria, indicating that growing methylotrophic
populations were preyed upon (Pérez et al., 2018).

122 Lytic virus populations linked to C1-hosts were analysed using metagenome
viral contigs (mVCs), predicted using established tools. Using contigs >10 kbp (Roux
124 et al., 2019) VirSorter (Roux et al., 2015) predicted 270 metagenome viral contigs
(mVCs) with a further 4 'likely' mVCs predicted uniquely by DeepVirFinder (Ren et al.,
126 2020) (see Supplementary Text), together representing 227 viral operational
taxonomic units (vOTUs) (Paez-Espino et al., 2017). Analysis of the normalised read
128 mapping for mVCs demonstrated that, as with the bacterial communities, active ¹³C-
enriched viral populations were reproducibly distinct between acidic and neutral pH
130 soils (Supplementary Fig. 3).

mVCs were linked to host bacteria using three different approaches: identifying
132 incorporation of viral DNA into spacers of bacterial CRISPR arrays, similarity of
homologous genes possessed by both host and virus contigs, and *k*-mer similarity
134 between potential host and virus contigs, the last approach being considered only
partially successful (see Supplementary Text). CRISPR arrays were identified in 3 of

136 23 MAGs, each associated with the genus *Methylocystis* or *Methylosinus* of the
Methylocystaceae (Fig. 1). In the acidic soil, complete CRISPR arrays of growing
138 methanotrophs were associated with two *Methylocystis* MAGs (MAG identifiers 5 and
6) sharing 79.2% average nucleotide identity (ANI) and likely representing different
140 species (Jain et al., 2018). A further six CRISPR arrays were identified in unbinned
bacterial contigs all possessing the same direct repeat (DR) sequence. These eight
142 arrays varied in size ranging from 9 to 114 DRs and contained a total of 432 spacers,
and were in the same size range of *Methylocystaceae* CRISPR arrays from sequenced
144 genomes (Supplementary Table 3). Comparison of spacer incorporation between
arrays revealed that these multiple closely-related populations had different histories
146 of viral interaction and subsequent spacer incorporation. Genome sequences from
¹³C-enriched viral populations were represented by seven mVCs and matched 29.5%
148 of spacers. In addition, 7.9% of spacers possessed a one nucleotide mismatch, all of
which represented a synonymous substitution, indicating that variation was the result
150 of mutations in viral genomes increasing their ability to evade CRISPR-CAS defense
systems or genetic variation in closely related viral populations. Only three pairs of
152 spacers were identical, with each pair member located on a different array. Variation
in virus host range was also observed, with 3 and 2 mVCs linked to only one or both
154 *Methylocystis* MAGs, respectively.

Surprisingly, a large number of spacers in individual *Methylocystis* CRISPR
156 arrays were linked to the same virus, with up to 31 being homologous to protospacer
sequences in one mVC. To provide support that these multiple spacers were derived
158 from *Methylocystis*-associated viruses, mVCs were examined for host-specific
conserved protospacer-adjacent motif (PAM) sequences (Mojica et al., 2009).
160 Consistent with the identification of genuine protospacers, 138 of linked 146 spacers
(i.e. all possessing ≤ 1 mismatch) had the conserved PAM sequence 'TTC' (target-
162 centric orientation) (Leenay and Beisel, 2017). The variation in spacer position
between the arrays revealed temporal differences in virus infections. For example, the
164 most recently integrated spacer in 3 of 8 different *Methylocystis* CRISPRs was derived
from a ¹³C-enriched virus represented by mVC_12213_cat2, suggesting the possibility
166 of incorporation occurring during the incubation of the experiment.

Further analysis of all CRISPR arrays (i.e. including those in unbinned contigs)
168 linked to ¹³C-enriched viruses revealed that the majority of viruses were associated
with methanotrophic populations (Fig. 2a). In total, 11 different variants were identified

170 (i.e. each having a unique DR sequence) with 9 linked to *Methylocystaceae* or
172 *Methylococcaceae* populations. DR sequences generally possessed high sequence
174 similarity to those in CRISPR arrays from cultivated strain genomes of the same family,
176 although only CRISPR array 6 had a DR sequence that was identical (Supplementary
178 Table 3). Individual DR variants were restricted to either pH 4.5 or 7.5 soil. Using 100%
180 sequence identity in searches between CRISPR spacer and mVC protospacer
182 sequences, 19 VirSorter-predicted mVCs were linked to all CRISPR array variants. In
addition, analysis of shorter mVCs ranging 5-10 kbp identified two additional linked
mVCs (mVC_08964_cat.3 (9.8 kbp) and mVC_28139_DVF (5.1 kbp)). One third of
CRISPR linked-mVCs were categorized at the lowest level of confidence (i.e.
category-3 by VirSorter (Roux et al., 2015) or 'possible' by DeepVirFinder (Ren et al.,
2020), suggesting that retaining only higher confidence contigs may exclude a
substantial proportion of *bona fide* methanotroph virus-derived contigs in
uncharacterised environments such as soil.

184 Analysis of tetranucleotide frequencies (TETRA) (Wang et al., 2017) clustered
the 21 mVCs into three groups that were associated with the *Methylocystaceae*,
186 *Methylococcaceae* and an unknown group (Fig. 2b). The majority of viruses infected
members of the *Methylocystaceae* family, with those infecting populations of the
188 *Methylocystis* and *Methylosinus* genera restricted to acidic and neutral pH soils,
respectively. TETRA correlation coefficients of all *Methylocystaceae*-linked viruses
190 were in the same range both within and between either genus, suggesting co-evolution
with their host rather than genetic drift and divergence was the primary mechanism for
192 defining specific associations with *Methylocystis* or *Methylosinus* strains.

Identification of homologous genes in mVCs that were shared with prokaryotic
194 genomes were always consistent with host-virus linkages established using spacer
sequences from MAG CRISPR arrays. Specifically, BLASTp searches of genes
196 present in the 9 mVCs linked to *Methylocystaceae* MAGs via CRISPR spacer
sequences all contained 'best hits' (identity >30%, e-value <10⁻⁵, bit score >50 and
198 query cover >70%) to a minimum of 5 homologues also found in *Methylocystaceae*
genomes. This was therefore used as a criterion for establishing host-virus linkages.
200 Sixty-three percent of mVCs contained a homologue that was linked to genomes of
known C1-utilising bacteria, with 35% linked specifically to populations from the
202 *Methylocystaceae*, *Methylococcaceae* or *Hyphomicrobiaceae* (Fig. 3a). While
analysis of bacterial homologues in mVCs identified the taxonomic family of the

204 assumed dominant host, they also indicated that individual viruses may infect hosts of
other families of the same taxonomic order, including those at other trophic levels.
206 Specifically, within the *Rhizobiales*, mVCs linked to *Methylocystaceae* also contained
homologues shared with *Bradyrhizobiaceae*, *Methylobacteriaceae* and *Rhizobiaceae*
208 (Fig. 3b), indicating that viruses of methanotrophs may also infect non-methanotrophic
methylophages that are active at the same time.

210 CH₄-derived C was also transferred to viruses of secondary or tertiary utilisers.
One group of mVCs were linked to methylophagous *Hyphomicrobiaceae* and a second
212 to a phylogenetically diverse range of nitrogen-fixing Rhizobia i.e. *Bradyrhizobiaceae*,
Phyllobacteriaceae and *Rhizobiaceae*. These lineages contain known methylophages,
214 methanol dehydrogenases have been identified in a range of rhizobial species (Huang
et al., 2019) and these mVCs also contained homologues found in the genomes of
216 nodulating *Methylobacterium* strains (Green and Ardley, 2018). Viruses of predatory
Bdellovibrio and *Myxococcales* bacteria were predicted, consistent with the recovery
218 of corresponding bacterial MAGs in ¹³C-enriched DNA. One mVC (20210-cat_2) was
linked to the genus *Bdellovibrio* (sharing 11 of 67 mVC genes) and three category-3
220 mVCs (i.e. possible viruses) was linked to *Myxococcales* populations, containing gene
homologous to four families within the order. The high isotopic labelling of both
222 heterotrophic predators (with no identifiable C1-utilising capability) and their viruses
indicate that the predators were feeding primarily on C1-utilisers, as carbon
224 incorporated by feeding on unlabelled bacteria would dilute the enrichment in
predators. As such, it also indicates that predatory bacteria may have preference for
226 preying upon growing populations rather than the non-C1-utilising majority.

Gene-sharing network analysis of mVCs with viruses in the NCBI RefSeq
228 database and other metagenome studies were analysed using vConTACT 2.0 (Jang
et al., 2019). Any linkages with RefSeq viruses typically had low scores (i.e. sharing a
230 low number of homologues) and were linked to viruses of hosts that were inconsistent
with our homologue-based predictions (Supplementary Table 4). No linkages were
232 observed with recently reported giant viruses of methanotrophs in freshwater lakes
(Chen et al., 2020). However, in a recent study of 197 metagenomes from Swedish
234 peatland soil, Emerson et al. (2018) identified 13 viruses linked to methanotrophs.
Intriguingly, 8 of these were linked in our viral gene-sharing network, with both studies
236 predicting *Methylocystaceae* hosts using different methods of host annotation
(Supplementary Text, Supplementary Fig. 5) and revealing the distribution of specific

238 *Methylocystaceae*-associated viral groups present in different geographical areas and
soil types. Analysis of gene-sharing networks of mVCs from this study indicated that
240 there were two distinct *Methylocystaceae*-linked viral clusters which also varied in their
distribution in both soils. Specifically, one cluster was associated with low pH only
242 whereas the second cluster contained viruses found in both pH 4.5 and 7.5 soils,
including those linked by CRISPR spacer sequences. Individual networks of
244 *Methylococcaceae*- and rhizobia-associated mVCs were also identified, typically
associated with one of the two soils of contrasting pH. Taxonomically-linked mVCs
246 with ≥ 5 homologues were consistently placed in networks with other mVCs containing
1-4 homologues from the same methylophilic families, confirming host linkage to a
248 larger number of mVCs.

mVCs contained 8,174 genes, with 49.6% (4,054) annotated representing 606
250 unique functions. Of these, genes encoding viral proteins accounted for 9.8% (397
genes) and included major capsid proteins, tail proteins, integrases, portal proteins
252 and terminases. Bacterial proteins used for viral replication accounted for 5.1% (206
genes). A number of metagenomic studies have demonstrated that viruses can
254 possess genes encoding sub-unit C of ammonia or particulate methane
monooxygenases as auxiliary metabolic genes (AMGs) (Chen et al., 2020; Ahlgren et
256 al., 2019) which are also typically found as isolated genes in genomes in addition to
being present in clusters or operons encoding A and B sub-units (Nicol and Schleper,
258 2006). In this study, one low confidence mVC (7.3 kbp, category-3) was identified as
containing an isolated *pmoC* gene that was phylogenetically related to growing
260 *Methylocystis* populations but was distinct from *pmoC* sequences found in viruses
associated with freshwater *Methylocystis* populations (Chen et al., 2020)
262 (Supplementary Fig. 6).

In summary, these results demonstrate that by following carbon flow, viruses
264 and hosts associated with a critical biogeochemical process can be identified at the
scale of individual populations, and currently active interactions at different trophic
266 levels examined within the highly complex soil environment. Type I and II
methanotrophs interact with evolutionarily distinct groups of viruses and the
268 composition of CRISPR arrays of *Methylocystaceae* reveal that they have a continual
dynamic interaction with individual viruses. Analysis of shared homologues in

270 individual viral genomes show that they may interact with host populations at different
271 trophic levels within a methane-fuelled network.

272

273 **Methods**

274 *Soil microcosms*

275 Triplicate soil samples were collected in February 2018 at 1 m intervals from the upper
276 10 cm of pH 4.5 and 7.5 soil sub-plots of a pH gradient maintained since 1961 and
277 under an 8-year crop rotation (SRUC, Craibstone Estate, Aberdeen, Scotland; UK grid
278 reference NJ872104) (Kemp et al., 1992). The crop at the time of sampling was
279 potatoes. Soil (podzol, sandy-loam texture) was sieved (2 mm mesh size) and
280 microcosms established in triplicate for each soil pH and isotope in 144 ml serum
281 bottles containing 14.30 g soil (10 g dry weight equivalent) with a 30% volumetric water
282 content, equivalent to ~60% water-filled pore space. Bottles were capped and
283 established with a 10% (v/v) $^{12}\text{C-CH}_4$ or $^{13}\text{C-CH}_4$ (Sigma-Aldrich) headspace (99%
284 atom enriched), re-opening every 10 days to maintain aerobic conditions before
285 sealing and re-establishing CH_4 headspace concentrations. Microcosms were
286 incubated at 25°C and destructively sampled after 30 days with soil archived
287 immediately at -20°C.

288

DNA-SIP

289 Genomic DNA was extracted from 0.5 g soil samples using a CTAB buffer
290 phenol:chloroform: isoamyl alcohol bead-beating protocol and subjected to isopycnic
291 centrifugation in CsCl gradients, recovery and purification as previously described
292 (Nicol and Prosser, 2011). Briefly, 6 μg of genomic DNA was added to 8 ml CsCl-Tris
293 EDTA solution (refractive index (RI) of 1.4010; buoyant density of 1.71 g ml^{-1}) in
294 polyallomer tubes before sealing and ultracentrifugation at 152,000 $\times g$ (50,000 rpm)
295 in a MLN80 rotor (Beckman-Coulter) for 72 h at 25°C. CsCl gradients were fractionated
296 into 350 μl aliquots using an in-house semi-automated fraction recovery system before
297 determining RI and recovering DNA. The relative abundance of bacterial 16S rRNA
298 and methanotrophic *pmoA* genes in genomic DNA distributed across the CsCl
299 gradients was determined by qPCR in a Corbett Rotor-Gene 6000 thermocycler
300 (Qiagen) using primer sets P1(341f)/P2(534r) (Muyzer et al., 1993) and A189F/A682R
301 (Holmes et al., 1995) respectively. Twenty-five μl reactions contained 12.5 μl 2X
302 QuantiFast SYBR Green Mix (Qiagen), 1 μM of each primer, 100 ng of T4 gene protein

304 32 (Thermo Fisher), 2 μ l of standard (10^8 - 10^2 copies of an amplicon-derived standard)
or 1/10 diluted DNA. Thermocycling conditions consisted of an initial denaturation step
306 of 15 min at 95°C for both assays followed by 30 cycles of 15 s at 94°C, 30 s at 60°C,
30 s at 72°C for the 16S rRNA gene assay or 60 s at 94°C, 60 s at 56°C, 60 s at 72°C
308 for the *pmoA* gene assay, followed by melt-curve analysis. All assays had an efficiency
between 93-97% with an r^2 value >0.99 . Genomic DNA from four fractions with a
310 buoyant density >1.732 g ml $^{-1}$ were then pooled for each ^{12}C - and ^{13}C -CH $_4$ -incubated
replicate for 16S rRNA gene amplicon sequence and metagenomic analysis.

312

Metagenome sequencing, assembly & annotation

314 Library preparation and sequencing was performed at the Joint genome Institute (JGI),
Berkeley, USA. Libraries were produced from fragmented DNA using KAPA
316 Biosystems Library Preparation Kits (Roche) and quantified using KAPA Biosystems
NGS library qPCR kits. Indexed samples were sequenced (2 x 150 bp) on the Illumina
318 NovaSeq platform with NovaSeq XP v1 reagent kits and a S4 flowcell. Raw reads
were processed with JGI's RQCFilter2 pipeline that utilised BBTools v38.51 (Bushnell,
320 2016). Reads containing adapter sequences were trimmed and those with ≥ 3 N bases
or ≤ 51 bp or $\leq 33\%$ of full-read length were removed along with PhiX sequences using
322 BBDuk, and reads mapped to human, cat, dog or mouse references at 95% identity
were removed using BBDuk. *De novo* contig assembly of the 100 - 196 million quality-
324 controlled reads per metagenome was performed using MetaSPAdes v3.13.0 (Nurk
et al., 2016). The 1 - 2 million contigs per metagenome were then concatenated
326 together, and contigs larger than 5 kbp were dereplicated at 99% average nucleotide
identity (ANI) using PSI-CD-HIT v4.6.1 (Fu et al., 2012) and binned using MetaWRAP
328 v1.2.1 (Uritskiy et al., 2018) (Supplementary Table 5). Bin completion and
contamination was determined by CheckM v1.0.12 (Parks et al., 2015). Taxonomic
330 annotation of contigs was performed using Kaiju (Menzel et al., 2016) with the NCBI
RefSeq database (Release 94; 25 June 2019) (O'Leary et al., 2016) and MAGs using
332 GTDB-Tk v0.3.2 (Chaumeil et al., 2019) with the Genome Taxonomy Database
(release 89, 21 June 2019) (Parks et al., 2018). Protein sequence annotation was
334 performed using InterProScan 5 (e-value $<10^{-5}$) (Jones et al., 2015). Pairwise ANI
comparison of MAGs was calculated using FastANI (Jain et al., 2019).

336

Amplicon sequencing and analysis

338 16S rRNA genes were amplified using primers 515F/806R (Walters et al., 2015)
followed by library preparation and sequencing on an Illumina MiSeq sequencer as
340 previously described (Finn et al., 2020). Reads with a quality score <20 and length <
100 bp were discarded using FASTX-Toolkit v0.0.13
342 (http://hannonlab.cshl.edu/fastx_toolkit/). High-quality reads were merged using
PANDAseq v2.11 (Masella et al., 2012), and denoising and chimera removal
344 performed with UNOISE3 (Edgar, 2016). Amplicon sequence variants (ASVs) were
annotated using the RDP classifier v2.11 (Wang et al., 2007). Non-metric
346 multidimensional scaling of Bray-Curtis dissimilarity derived from the relative
abundance of ASVs was performed with the metaMDS function in the vegan package
348 (Oksanen et al., 2019) in R v3.6.0.

350 *Virus prediction*

Metagenomic viral contigs (mVCs) were predicted from 9,190 contigs >10 kbp using
352 VirSorter (Roux et al., 2015), retaining non-prophage category-1, -2 or -3 mVCs,
representing “most confident”, “likely” and “possible”. DeepVirFinder (Ren et al., 2020)
354 was also used to predict mVCs from contigs >10 kbp, with those with a p -value <0.05
and a score ≥ 0.9 or ≥ 0.7 , representing “confident” and “possible”, respectively
356 (Supplementary Table 6). The relative abundance of each mVC in the six
metagenomes was determined using the MetaWRAP-Quant_bins module (Uritskiy et
358 al., 2018) and a heatmap produced using the heatmaply package in R v3.6.0.

360 *Virus-host linkage*

CRISPR arrays within MAGs and unbinned contigs were identified using the CRISPR
362 Recognition Tool v1.2 (Bland et al., 2007) (Supplementary Table 7). DR and spacer
sequences were extracted before performing 100% identity searches against positive
364 and negative strands to identify MAGs or contigs with direct repeats and the viral origin
of spacers using Seqkit commands (Shen et al., 2016). After identification of matched
366 spacer sequences in mVCs, 10 nucleotides before and after the spacer sequence
were extracted to identify associated host-specific PAM sequences. Conserved and
368 variant PAM sequences were manually identified. Correlation coefficients of pairwise
comparison of the tetra-nucleotide frequencies (TETRA) between unique CRISPR-
370 associated mVCs were calculated using Python package pyani v0.2.10 (Pritchard et
al., 2016). To identify homologous genes shared between CRISPR-linked viruses and

372 hosts, gene prediction was performed using Prodigal v2.6.3 (Hyatt et al., 2010) with
the -p meta option followed by protein alignment with Blastp (identity >30%, e-value <
374 10^{-5} , bit score >50 and query cover >70%) and protein sequence annotation using
InterProScan 5 (e-value < 10^{-5}). Gene homology between all mVCs and prokaryotes in
376 the NCBI nr database was determined using Diamond Blastp (e-value < 10^{-5}) (Buchfink
et al., 2015). Virus-host prediction using k-mer frequencies was performed with WISH
378 v1.0 (p-values <0.05) (Galiez et al., 2017). Networks based on shared gene content
was constructed using vConTACT 2.0 (Jang et al., 2019) with the NCBI RefSeq
380 database (Release 94; 25 June 2019).

382 *Phylogenetic analysis of PmoC and PxmC protein sequences*

Maximum likelihood analysis of inferred protein sequences of membrane-bound
384 monooxygenase C sub-units from methanotroph MAGs and reference sequences
(Supplementary Table 8) was performed on unambiguously aligned sequences using
386 PhyML 3.0 (Guindon et al., 2010) with automatic model selection (LG substitution,
gamma distribution (0.06) and proportion of invariable sites (0.087) estimated).
388 Bootstrap support was calculated from 100 replicates.

390 **Data availability**

Metagenome sequence reads are deposited under NCBI BioProject accession
392 numbers PRJNA621430 - PRJNA621435. Metagenome draft assemblies are
accessible through the JGI Genome Portal (DOI: 10.25585/1487501). Amplicon
394 sequence data is deposited in the NCBI Sequence Read Archive with BioProject
accession number PRJNA676099.

396

Acknowledgments

398 The sequencing data were generated under JGI Community Science Program
proposal 503702 awarded to GWN and CH. The work conducted by the U.S.
400 Department of Energy Joint Genome Institute, a DOE Office of Science User Facility,
is supported by the Office of Science of the U.S. Department of Energy under Contract
402 No. DE-AC02-05CH11231. This work was funded by an AXA Research Chair awarded
to GWN and a France-Berkeley Fund grant (2018-2019) awarded to GWN and MKF.
404 The pH gradient experiment is funded through the Scottish Government RESAS 2016-

2021 programme. The authors would like to thank Dr. Joanne Emerson for valuable
406 discussion.

408 **Author contributions**

The research program was conceived by and funded from grants awarded to GWN,
410 CH and MF. SL, CH and GWN designed the experiment and wrote the manuscript. SL
performed experiments and analyses. ES, AN and MF advised on bioinformatic
412 approaches, discussed data and commented on the manuscript. RW coordinated soil
sampling and commented on the manuscript. All authors approved the manuscript.

414 **References**

416 Ahlgren, N.A., Fuchsman, C.A., Rocap, G. & Fuhrman, J.A. Discovery of several
418 novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that
encode *amoC* nitrification genes. *ISME J.* 13, 618-631 (2019).

420 Angel, R., Claus, P., & Conrad, R. Methanogenic archaea are globally ubiquitous in
aerated soils and become active under wet anoxic conditions. *ISME J.* 6, 847-862
422 (2012).

Bland, C. et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of
424 clustered regularly interspaced palindromic repeats. *BMC Bioinform.* 8, 209 (2007).

Bowers, R.M. et al. Minimum information about a single amplified genome (MISAG)
426 and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat.*
Biotech. 35, 725-731 (2017).

428 Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using
DIAMOND. *Nat. Methods* 12, 59-60 (2015).

430 Bushnell, B. BBTools software package. <http://sourceforge.net/projects/bbmap>
(2016).

432 Carini, P., Marsden, P.J., Leff, J.W., Morgan, E.E., Strickland, M.S. & Fierer N. Relic
DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat.*
434 *Microbiol.* 2, 16242 (2017).

Chaumeil, P-A., Mussig, A.J., Hugenholtz, P. & Parks, D.H. GTDB-Tk: a toolkit to
436 classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36, 1925-
1927 (2019).

- 438 Chen, L.-X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat. Microbiol.* In press (2020).
- 440 Chistoserdova, L., Kalyuzhnaya, M.G. & Lidstrom, M.E. The expanding world of methylotrophic metabolism. *Annu. Rev. Microbiol.* **63**, 477–499 (2009).
- 442 Curry, C.L. Modeling the soil consumption of atmospheric methane at the global scale, *Global Biogeochem. Cycles* **21**, GB4012 (2007).
- 444 Edgar, R.C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. Preprint at <https://doi.org/10.1101/081257> (2016).
- 446 Emerson, J.B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870-880 (2018).
- 448 Emerson, J.B. Soil viruses: A new hope. *mSystems* **4**, e00120-19 (2019).
- Finn, D.R., Lee, S., Lazén, M.B., Nicol, G.W. & Hazard, C. Cropping systems that
450 improve richness convey greater resistance and resilience to soil fungal, relative to
prokaryote, communities. Preprint at <https://doi.org/10.1101/2020.03.15.992560>
452 (2020).
- Frossard, A., Hammes, F., & Gessner, M.O. Flow cytometric assessment of bacterial
454 abundance in soils, sediments and sludge. *Front. Microbiol.* **7**, 903 (2016).
- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
456 generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
- Galiez, C, Siebert, M., Enault, F., Vincent, J. & Söding, J. WisH: who is the host?
458 Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**,
3113-3114 (2017).
- 460 Gómez, P. & Buckling. Bacteria-phage antagonistic coevolution in soil. *Science* **332**,
106-109 (2011).
- 462 Green, P.N and Ardley, J.K. Review of the genus *Methylobacterium* and closely
related organisms: a proposal that some *Methylobacterium* species be reclassified into
464 a new genus, *Methylorubrum* gen. nov. *Int. J. Syst. Evol. Microbiol.* **68**, 2727-2748
(2018).
- 466 Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood
phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307-21 (2010).

- 468 Holmes, A.J., Costello, A., Lidstrom, M.E. & Murrell, J.C. Evidence that particulate
methane monooxygenase and ammonia monooxygenase may be evolutionarily
470 related. *FEMS Microbiol. Lett.* **132**, 203-208 (1995).
- Huang, J. et al. Rare earth element alcohol dehydrogenases widely occur among
472 globally distributed, numerically abundant and environmentally important microbes.
ISME J. **13**, 2005-2017 (2019).
- 474 Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site
identification. *BMC Bioinform.* **11**, 119 (2010).
- 476 Ignacio-Espinoza, J. C., Ahlgren, N.A. & Fuhrman, J.A. Long-term stability and Red
Queen-like strain dynamics in marine viruses. *Nat. Microbiol.* **5**, 265-271 (2020).
- 478 Jain, C., Rodriguez-R, L.M., Phillippy, A.M. & Konstantinidis, K.T. High throughput ANI
analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Comm.*
480 **9**, 5114 (2018).
- Jang, H.B. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is
482 enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632-639 (2019).
- Jones, P. et al. InterProScan 5: genome-scale protein function classification.
484 *Bioinformatics* **30**, 1236-1240 (2015).
- Kemp, J.S., Paterson, E., Gammack, S.M., Cresser, M.S. & Killham, K. Leaching of
486 genetically modified *Pseudomonas fluorescens* through organic soils: influence of
temperature, soil pH, and roots. *Biol. Fert. Soils* **13**, 218–224 (1992).
- 488 Kirschke, S., et al. Three decades of global methane sources and sinks. *Nat. Geosci.*
6, 813–823 (2013).
- 490 Knief, C. Diversity and habitat preferences of cultivated and uncultivated aerobic
methanotrophic bacteria evaluated based on *pmoA* as molecular marker. *Front.*
492 *Microb.* **6**, 1346 (2015).
- Le Mer, J. & Roger, P. Production, oxidation, emission and consumption of methane
494 by soils: a review. *Eur. J. Soil Biol.* **37**, 25–50 (2001).
- Leenay, R.T. & Beisel, C.L. Deciphering, communicating, and engineering the
496 CRISPR PAM. *J. Mol. Biol.* **429**, 177-191 (2017).

- 498 Lyu, Z., Shao, N., Akinyemi, T. & Whitman, W.B. Methanogenesis. *Curr. Biol.* **28**, 727–
732 (2018).
- 500 Masella, A.P., Bartram, A.k., Truszkowski, J.M., Brown, D.G. & Neufeld, J.D.
PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinform.* **13**, 31
(2012).
- 502 Menzel, P., Ng, K.L. & Krogh, A. Fast and sensitive taxonomic classification for
metagenomics with Kaiju. *Nat. Comm.* **7**, 11257 (2016).
- 504 Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. Short motif
sequences determine the targets of the prokaryotic CRISPR defence system. *Microb.*
506 **155**, 733-740 (2009).
- 508 Muyzer, G., De Waal, E.C. & Uitterlinden, A.G. Profiling of complex microbial
populations by denaturing gradient gel electrophoresis analysis of polymerase chain
reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* **59**, 695-700
510 (1993).
- 512 Nicol, G.W. & Prosser, J.I. Strategies to determine diversity, growth and activity of
ammonia oxidising archaea in soil. *Meth. Enzymol.* **496**, 3-34 (2011).
- 514 Nicol, G.W. & Schleper C. Ammonia-oxidising Crenarchaeota: important players in the
nitrogen cycle? *Trends Microbiol.* **14**, 207–212 (2006).
- 516 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new
versatile metagenomic assembler. *Genome Res.* **27**, 824-834 (2017).
- 518 O’Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status,
taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-D745
(2016).
- 520 Oksanen, J. et al. vegan: Community Ecology Package. [https://CRAN.R-](https://CRAN.R-project.org/package=vegan)
project.org/package=vegan (2019).
- 522 Paez-Espino, D., Pavlopoulos, G.A., Ivanova, N.N. & Kyrpides, N.C. Nontargeted virus
sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.*
524 **12**, 1673 (2017).

- 526 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
assessing the quality of microbial genomes recovered from isolates, single cells, and
metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 528 Parks, D.H. et al. A standardized bacterial taxonomy based on genome phylogeny
substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996-1004 (2018).
- 530 Pérez, J. Moraleda-Muñoz, A., Marcos-Torres, F.J. & Muñoz-Dorado, J. Bacterial
predation: 75 years and counting! *Environ. Microbiol.* **18**, 766–779 (2018).
- 532 Pratscher, J., Vollmers, J., Wiegand, S., Dumont, M.G., & Kaster, A.-K. Unravelling
the identity, metabolic potential and global biogeography of the atmospheric methane-
534 oxidizing Upland Soil Cluster α . *Environ. Microbiol.* **20**, 1016-1029 (2018).
- Pasulka, A.L. et al. Interrogating marine virus-host interactions and elemental transfer
536 with BONCAT and nanoSIMS-based methods. *Environ. Microbiol.* **20**, 671-692 (2018).
- Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G. & Toth, I.K. Genomics and
538 taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens.
Anal. Methods **8**, 12-24 (2016).
- 540 Radajewski, S., et al. Identification of active methylotroph populations in an acidic
forest soil by stable-isotope probing. *Microbiol.* **148**, 2331–2342 (2002).
- 542 Ren, J. et al. Identifying viruses from metagenomic data by deep learning. *Quant. Biol.*
8, 64-77 (2020).
- 544 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
microbial genomic data. *PeerJ* **3**, e985 (2015).
- 546 Roux, S. et al. Minimum information about an uncultivated virus genome (MIUViG).
Nat. Biotechnol. **37**, 29–37 (2019).
- 548 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A cross-platform and ultrafast toolkit for
FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
- 550 Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev.*
Microbiol. **5**, 801–812 (2007).
- 552 Uritskiy, G.V., DiRuggiero, J. & Taylor, J. MetaWRAP- a flexible pipeline for genome-
resolved metagenomic data analysis. *Microbiome* **15**, 158 (2018).

554 Walters, W. et al. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal
transcribed spacer marker gene primers for microbial community surveys. *mSystems*
556 **1**, e00009-15 (2015).

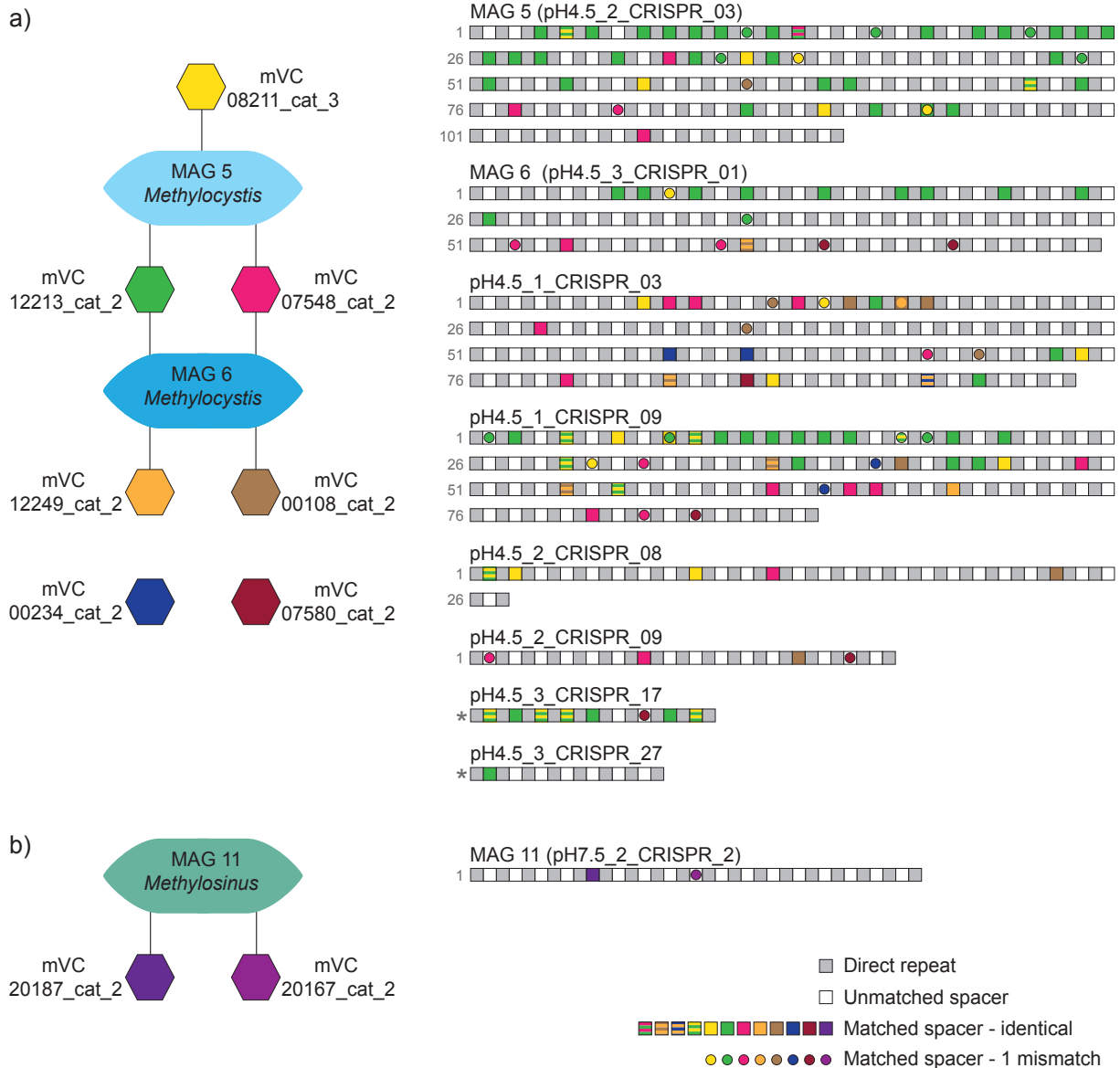
Williamson, K.E., Fuhrmann, J.J., Wommack, K.E., & Radosevich, M. Viruses in soil
558 ecosystems: an unknown quantity within an unexplored territory. *Annu. Rev. Virol.* **4**,
201-219 (2017).

560 Wang, J. et al. Genomic sequence of '*Candidatus Liberibacter solanacearum*'
haplotype C and its comparison with haplotype A and B genomes. *PLoS One* **12**,
562 e0171531 (2017).

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naïve bayesian classifier for rapid
564 assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ.*
Microbiol. **73**, 5261-5267 (2007).

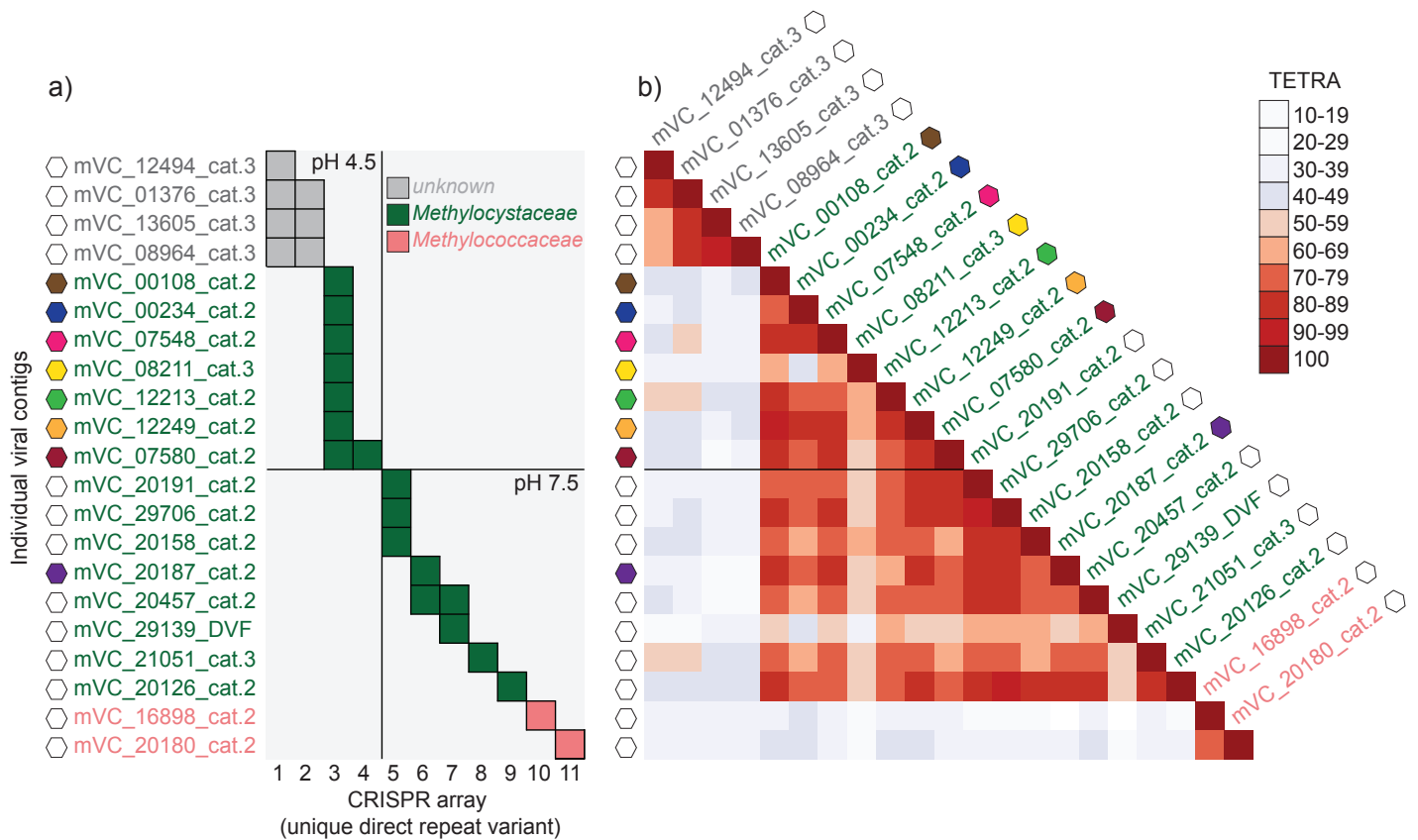
566 Zhao, J., Cai, Y. & Jia, Z. The pH-based ecological coherence of active canonical
methanotrophs in paddy soils. *Biogeosciences* **17**, 1451–1462 (2020).

568



560

Fig. 1. Linkage of active ^{13}C -enriched viruses to *Methylocystaceae* populations in soil
 562 by comparison of spacer sequences in CRISPR arrays. a) Distribution of spacers from
 564 8 mVCs in *Methylocystis* CRISPR arrays (MAGs 5, 6 and six unbinned contigs). b)
 566 CRISPR array of *Methylosinus* MAG 11 containing spacers linked to two mVCs.
 568 CRISPR array names describe the individual soil microcosm that the contig was
 570 recovered from. DRs for complete arrays are numbered (in grey), with the spacer after
 DR 1 being the most recently incorporated. Two partial arrays are denoted with an *.
 Spacers with 100% identity or 1 mismatch to sequences in mVCs are represented by
 colour-coded squares and circles, respectively, with stripes representing sequences
 found in two different mVCs.



572

Fig. 2. Linkages of ^{13}C -enriched viruses to CRISPR arrays in pH 4.5 and 7.5 soil. a)

574

Presence of spacers from 21 mVCs in 11 different CRISPR array variants (unique DR sequence). Taxonomic affiliation of CRISPR arrays to host families was determined

576

by phylogenomic analysis of affiliated MAGs (3, 6) or unbinned contigs (2-5, 7-9), or inferred from shared homologues between linked mVCs and bacterial genomes (10,

578

11). All mVCs were >10 kbp except mVC_08964_cat.3 (9.8 kbp) and mVC_28139_DVF (5.1 kbp). These two mVCs were also the only two predicted using

580

DeepVirFinder, with calculated probabilities describing 'likely' and 'probable' viruses, respectively. b) TETRA correlation coefficients between 21 CRISPR-linked mVCs.

582

Colour-coded hexagon symbols denote linkage to MAG-associated CRISPR arrays as per Fig. 1.

584

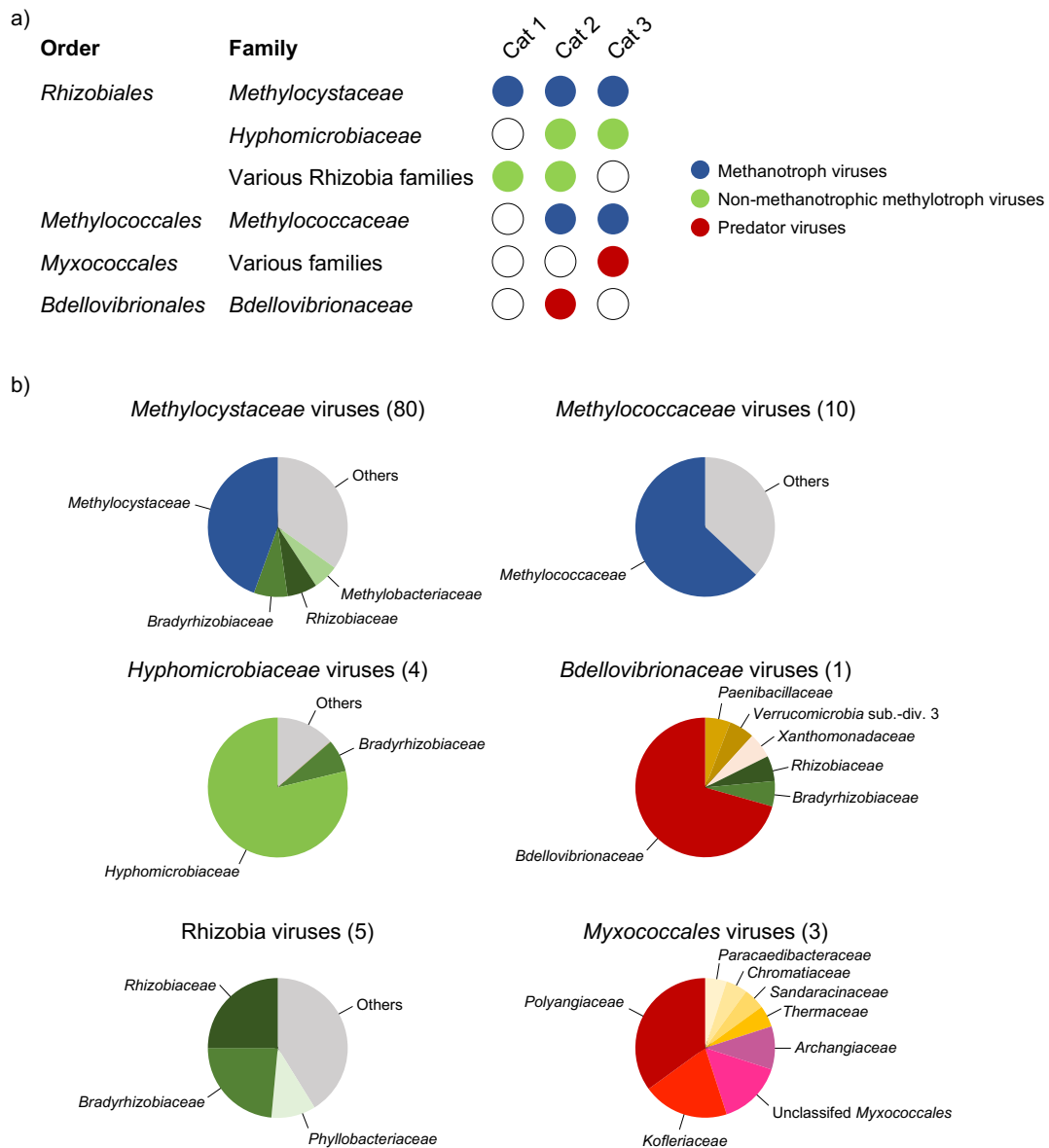
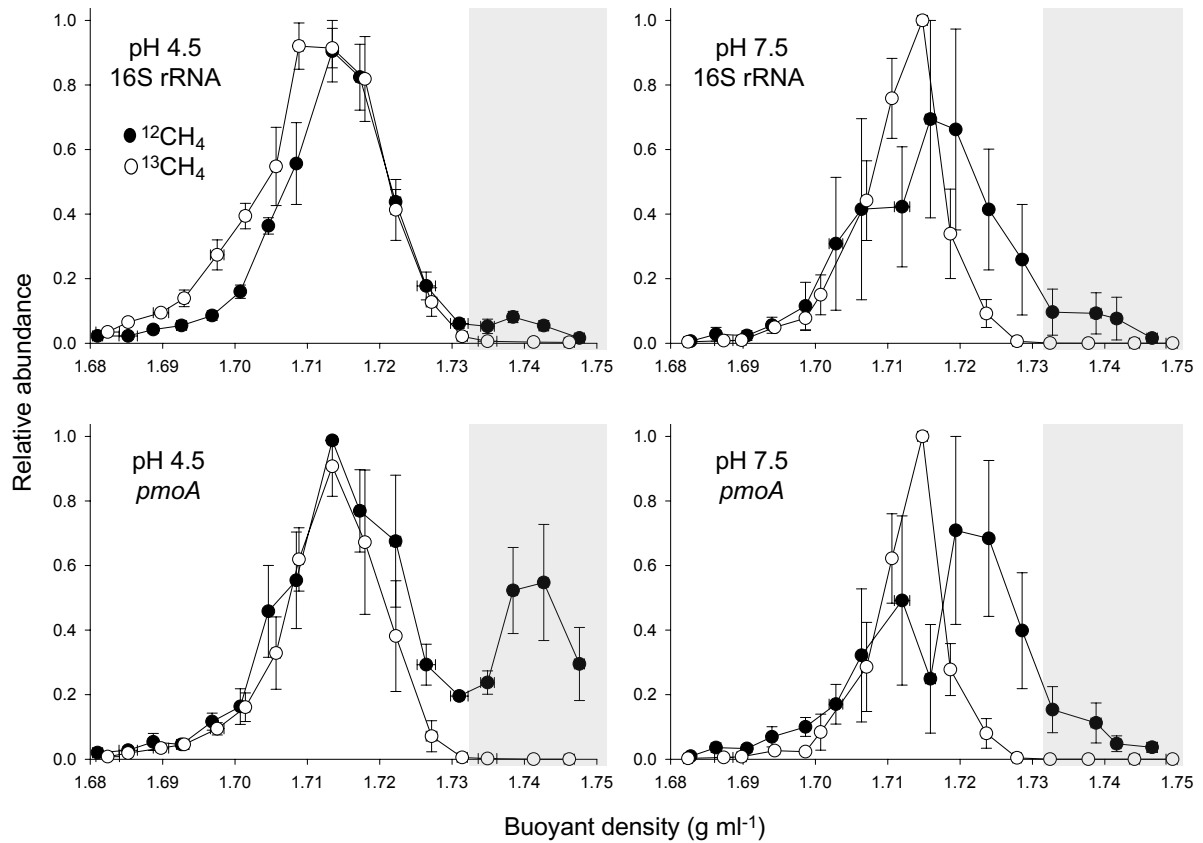


Fig. 3. Linkage of ^{13}C -enriched viruses to methanotrophic, methylotrophic and predator bacterial host populations through identification of shared homologous genes. a) Association of viruses with different bacterial families and functional groups inferred from the presence of ≥ 5 shared homologous genes in category-1, -2 and -3 VirSorter-predicted mVCs. b) Proportion of homologues in methanotroph, non-methanotrophic methylotroph or predator viruses linked to individual bacterial families. Each chart summarises those mVCs that all contain ≥ 5 homologues to one family (number of mVCs given in parentheses) but with other taxonomic linkages also given. 'Other' describes the proportion found in families each represented by less than $< 5\%$ of homologues or those not annotated to the family level.



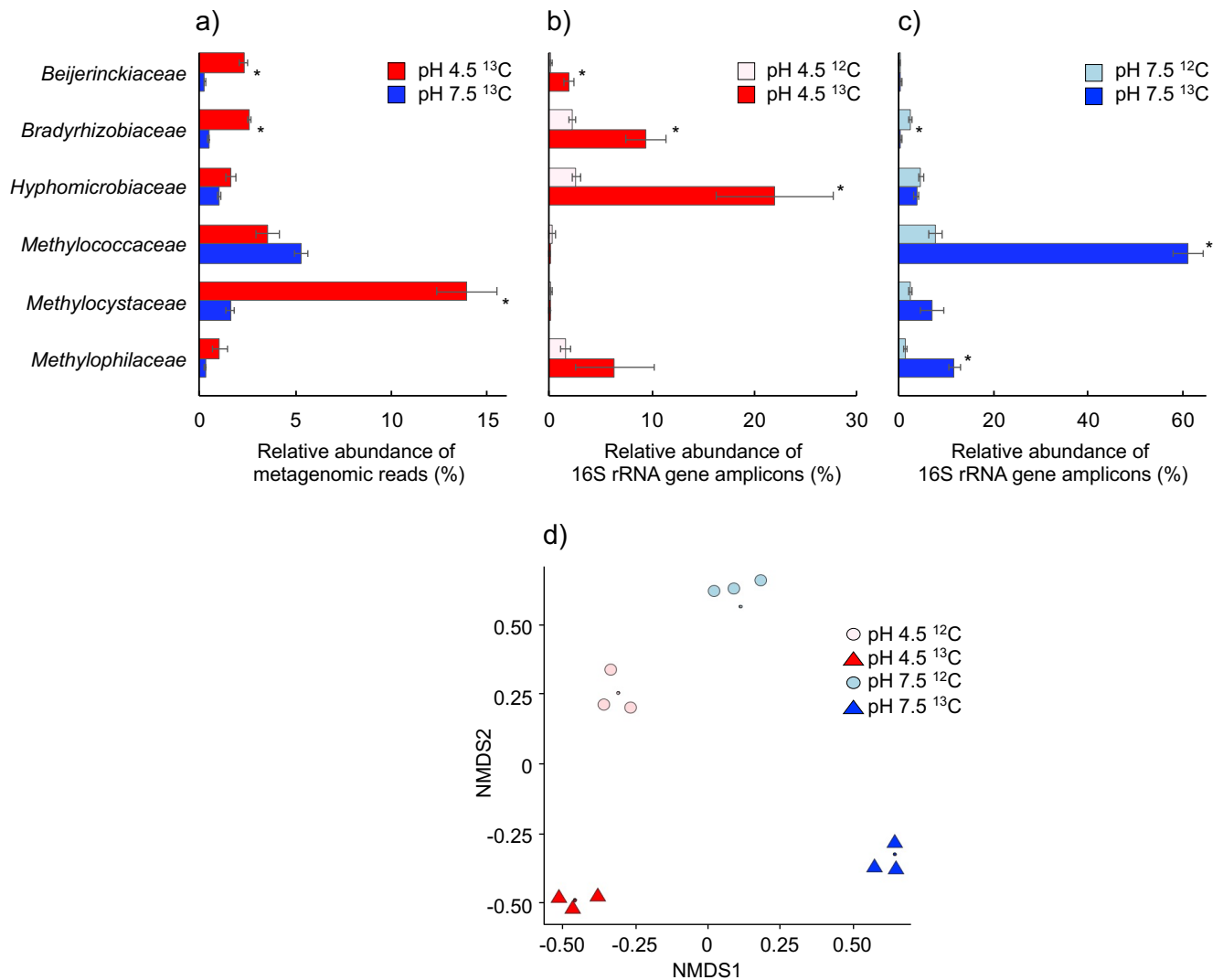
596

598

600

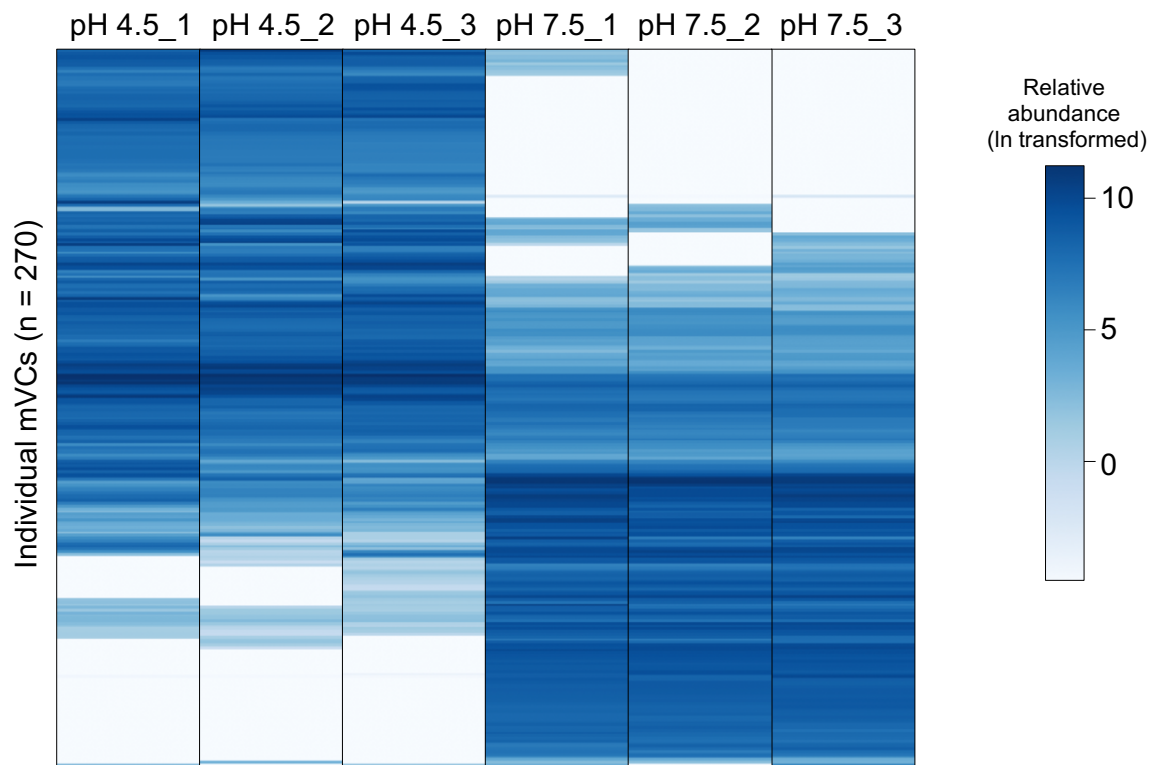
602 **Supplementary Fig. 1.** Buoyant density distribution of genomic DNA from total
603 bacterial 16S rRNA genes and methanotroph communities possessing particulate
604 methane monooxygenase sub-unit A (*pmoA*) genes after isopycnic centrifugation in
605 CsCl gradients. Genomic DNA was extracted from triplicate pH 4.5 and 7.5 soil
606 microcosms incubated with a 10% ¹²C- or ¹³C-CH₄ headspace. Vertical error bars are
607 the standard error of the mean relative abundance and horizontal bars (mostly smaller
608 than the symbol size) the standard error of the mean buoyant density of individual
609 fractions from three independent CsCl gradients, each representing an individual
610 microcosm. The four fractions with the highest buoyant density (highlighted by grey
611 area) were pooled for each replicate microcosm.

612



614 **Supplementary Fig. 2.** Taxonomic affiliation of metagenome reads and 16S rRNA
616 gene amplified sequence variants (ASVs) derived from high buoyant density DNA from
618 triplicate pH 4.5 and pH 7.5 soil microcosms after incubation with ¹²C- or ¹³C-CH₄. a)
618 Relative abundance of metagenome sequences mapped to contigs of families that
618 recruited ≥1% reads in either soil. Reads were mapped to annotated contigs ≥5 kbp
618 from ¹³C-incubated microcosms only. b) and c) Relative abundance of 16S rRNA gene
620 ASVs derived from the six dominant families in ¹²C- or ¹³C-CH₄ incubations of pH 4.5
620 and 7.5 soil, respectively. d) Non-metric multidimensional scaling of Bray-Curtis
622 dissimilarities derived from the relative abundance of annotated 16S rRNA gene
622 ASVs. Due to the close overlap of replicates (small symbols), samples were resolved
624 for visualisation using a jitter function (large symbols). Significant differences between
624 samples are indicated with * ($p < 0.05$, two-sample Student's t-test or Welsch's t-test
626 when variances were not homogenous).

628

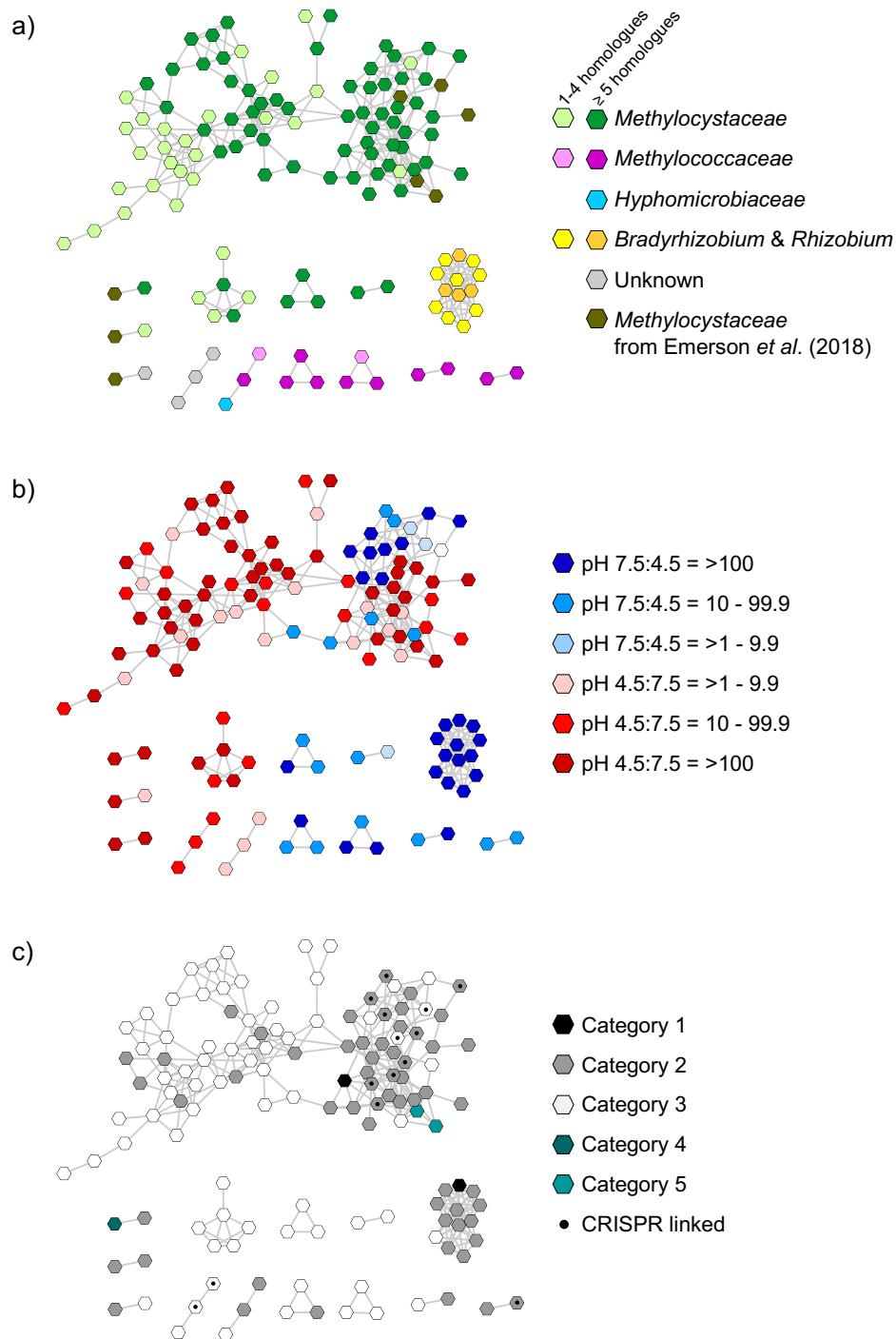


630

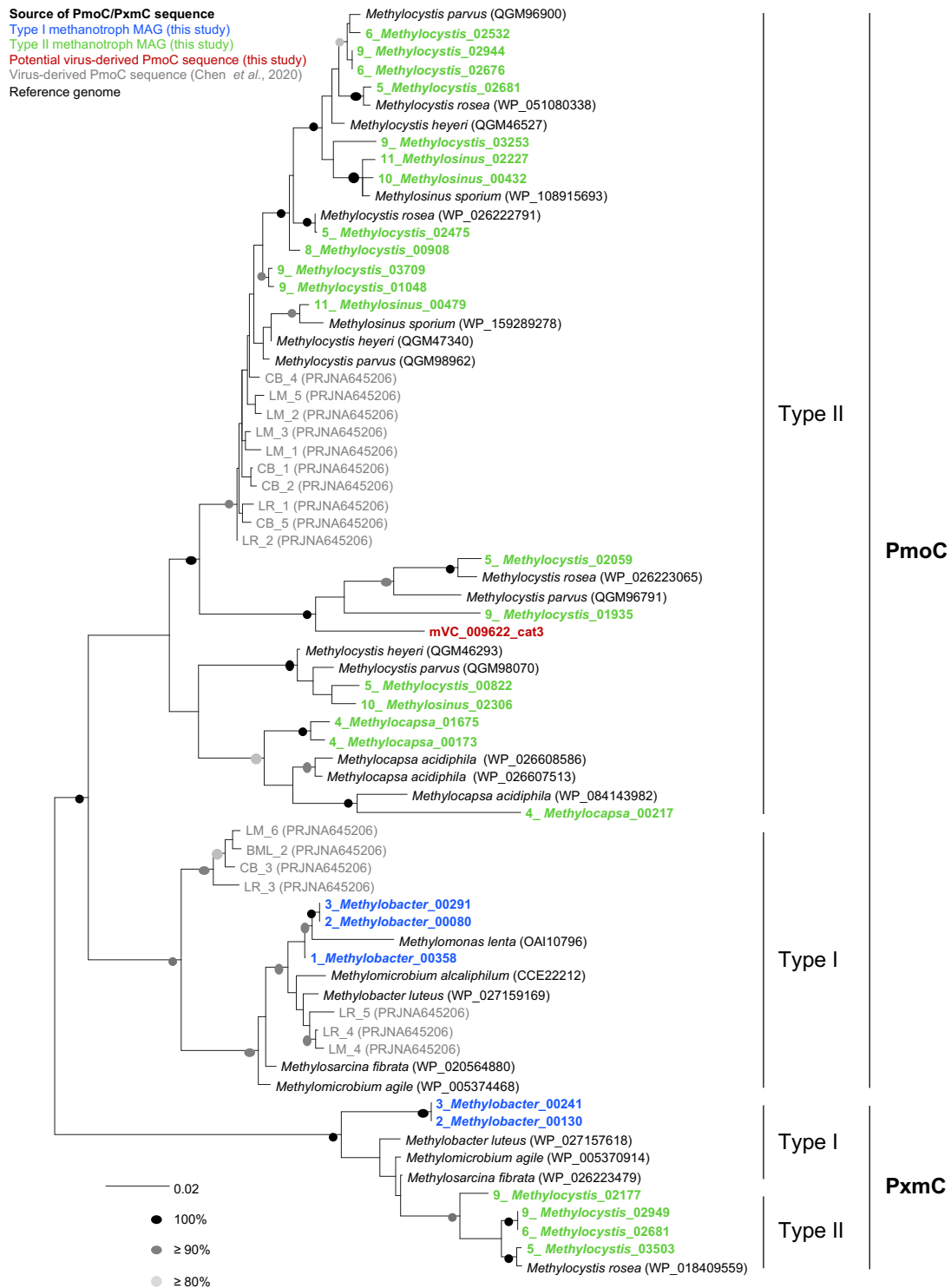
632

Supplementary Fig. 3. Heat-map displaying the relative abundance of 270 mVCs >10 kbp in length from ¹³C-enriched viral DNA derived from triplicate pH 4.5 and pH 7.5 soil microcosms. The values of normalised relative abundance are presented as reads per kbp after ln transformation.

636



638 **Supplementary Fig. 4.** Gene sharing network analysis of mVCs (one representative
640 per vOTU) from ^{13}C -enriched viruses in pH 4.5 and 7.5 soils. a) Taxonomic affiliation
of hosts predicted by homologue analysis, with mVCs containing ≥ 5 or < 5 linked
642 homologues highlighted. Eight mVCs from peatland soils linked to the
Methylocystaceae (Emerson *et al.*, 2018) are also shown. b) Distribution of mVCs in
644 pH 4.5 and 7.5 soil determined by the mean ratio of normalised relative abundance
from triplicate samples. Seven mVCs from peatland were only found in soils with pH
646 ≤ 4.7 . c) VirSorter category prediction and linkage to CRISPR arrays via spacer
sequence analysis.



Supplementary Fig. 5. Maximum likelihood phylogenetic tree of derived amino acid sequences of PmoC and PxmC found in three type I and seven type II methanotroph MAGs and one potential viral-derived sequence. MAG-derived sequences are described by MAG number, genus and contig identifier (see Supplementary Table 8). Sequences in reference methanotroph genomes and freshwater-derived viruses (Chen *et al.*, 2020) were included with NCBI accession numbers given in parenthesis. Circles at nodes describe percentage bootstrap support from 100 replicates and the scale bar represents 0.02 changes per amino acid position.

Supplementary text

656 *Comparison of 16S rRNA gene amplicon and metagenomic libraries of high buoyant*
657 *density DNA from ¹²C and ¹³C-CH₄ incubations*

658 The relative abundance of annotated contigs belonging to different families in replicate
659 metagenomic libraries was reproducible and distinct between acidic and neutral pH
660 soils (Supplementary Figure 2a). Twenty of the 23 medium and high-quality MAGs
661 recovered in ¹³C-derived metagenomic libraries were from methanotrophic or non-
662 methanotrophic methylotrophic populations and therefore consistent with targeting a
663 methane-fuelled community using stable isotope probing (Supplementary Table 2).
664 Nevertheless, comparison with equivalent ¹²C incubations was performed as per
665 standard practice with DNA-SIP experiments¹. Genomic DNA was recovered and
666 purified from high buoyant density fractions (>1.732 g ml⁻¹) from triplicate microcosms
667 of both ¹²C- and ¹³C-CH₄ incubations. While recovered DNA from ¹²C incubations was
668 considered too low for metagenome sequencing, with DNA concentrations below the
669 limit of detection in some fractions, PCR amplification enabled characterisation of 16S
670 rRNA gene-based community structures and comparison to the equivalent fractions
671 from ¹³C incubations.

672 Six taxonomic families were represented by contigs to which a minimum of ≥1%
673 of reads were mapped from metagenome analysis of ¹³C-enriched DNA in at least one
674 soil. These families represented 39.8% and 83.7% of 16S rRNA gene ASVs in pH 4.5
675 and 7.5 samples, respectively, compared to 6.9 and 18.3% in the equivalent DNA
676 fractions from ¹²C-incubations (Supplementary Fig. 2). Annotated metagenome and
677 16S rRNA amplicon libraries generated from the same ¹³C-enriched DNA contained
678 representatives of the same C1-utilising groups, although substantial differences were
679 observed in relative abundance. For example, while the *Methylococcaceae* was the
680 dominant family in pH 7.5 ¹³C-enriched DNA using both approaches, it represented
681 5.3% (±0.3% s.e.) and 61.1% (±3.2% s.e.) of annotated contigs and 16S rRNA gene
682 ASVs, respectively. In ¹³C-enriched pH 4.5 DNA, the dominant family in the
683 metagenomic libraries was the *Methylocystaceae*, representing 13.9% (±1.6% s.e.) of
684 annotated contigs but only 0.02% (±0.006% s.e.) in the amplicon libraries, where the
685 non-methanotrophic methylotrophic *Hyphomicrobiaceae* was the most abundant at
686 22.0% (±5.7% s.e.). Observed differences were therefore likely due to a combination
687 of variation in genome size and 16S rRNA gene copy number between strains in

688 addition to the range of biases associated with different marker-gene and
metagenomic sequencing approaches². Nevertheless, the overall community
690 composition determined by 16S rRNA gene amplicon libraries were highly
reproducible and clearly distinct between ¹²C and ¹³C incubations, confirming that
692 communities analysed in metagenomic libraries were enriched in methane-derived ¹³C
(Supplementary Fig. 2c).

694

Predicted C1 metabolism in metagenome assembled genomes

696 Aerobic methylotrophic organisms utilise C1 compounds such as methane or
methanol for both carbon and energy requirements³. Methanotrophs are one group of
698 methylotrophs that oxidise methane to formaldehyde via methanol, which is either
assimilated for generating biomass or oxidised through to CO₂ to obtain energy and
700 reductant. Non-methane oxidising methylotrophs, lacking methane monooxygenase,
utilise methanol produced from other sources including that excreted from
702 methanotrophs. Methane can therefore be directly and indirectly utilised by
methylotrophic populations in natural communities.

704 To determine whether MAGs represented methanotrophic or non-
methanotrophic methylotrophs, the presence of genes encoding methane
706 monooxygenase (MMO), methanol dehydrogenases (MDH) and formate
dehydrogenases (FDH) was determined after predicted protein sequence annotation.
708 Taxonomic assignment and methylotrophic characterisation were consistent with
known traits of *Methylobacter*, *Methylocapsa*, *Methylosinus* and *Methylocystis* strains,
710 all of which possessed particulate methane monooxygenase (pMMO) and with the
three *Methylosinus* MAGs also possessing soluble methane monooxygenase
712 (sMMO). Eight MAGS lacked genes encoding an MMO but possessed MDH and FDH
confirming methylotrophic capability. This included representatives of previously
714 recognised non-methanotrophic methylotrophs including *Gemmatimonadales*,
Hyphomicrobium, *Methylophilaceae*, *Methyloceanibacter* plus MAGs representative of
716 the genera *Rudea* and *Herminiimonas*. One MAG belonged to the class *Kiritimatiellae*
of the Verrucomicrobiota. While this phylum contains known methylotrophs, no
718 pathways for C1 metabolism were identified, potentially due to the low (51.5%)
estimated completeness.

720

722 *Comparison of VirSorter and DeepVirFinder in predicting virus-associated metagenome contigs*

Using assembled contigs >10 kbp, metagenomic viral contigs (mVCs) were predicted using two established tools. VirSorter⁴ uses a database of viral genes plus analysis of virus-like motifs, and DeepVirFinder⁵ uses a *k*-mer based alignment-free approach, using viral genomes to train the prediction model. Both approaches provide different levels of confidence. VirSorter categories 1, 2 and 3 represent ‘most confident’, ‘likely’ and ‘possible’ virus predictions, respectively, with categories 4, 5 and 6 the equivalent for proviruses. Based on probability values, DeepVirFinder virus predictions can be considered likely (≥ 0.9 , *p*-value <0.05) and probable (≥ 0.7 , *p*-value <0.05)⁶. We considered the matching of CRISPR array spacers with virus protospacers as the most confident method of confirming a viral origin for an individual contig and facilitated comparison of the success of the two prediction tools. Of the 21 mVCs linked using CRISPR spacer analysis, 19 were predicted using VirSorter only, 1 was predicted using DeepVirFinder only and 1 predicted using both, the latter two being <10 kbp and identified after further analysis of 5-10 kbp contigs. Of the 270 contigs predicted as category 1, 2 or 3 mVCs by VirSorter, 49 were also predicted by DeepVirFinder, which uniquely identified a further 41 (of which only 4 were ‘likely’ viruses). These results indicate that substantially more soil virus genomes are required for training datasets using an alignment-free approach.

742 *Evaluation of k-mer analysis for identifying host-virus linkages*

The matching of protospacers in mVCs to spacers in CRISPR arrays with 100% identity was also used to validate criteria for linkages via a ‘best hit’ homologue approach, specifically the sharing of a minimum of five homologues to the same taxonomic family. These two approaches were then compared to linkages predicted using a *k*-mer based analysis with the tool WIsH⁷, which involves the comparison and prediction of linkage of previously defined host- or virus-derived contigs on the basis of *k*-mer frequency analysis. Seven of the 21 mVCs predicted to hosts via CRISPR spacer analysis were linked to a host using WIsH and a probability score ≤ 0.05 (Supplementary Table 9), all of which were consistent at the family level. Of the 103 mVCs linked in vConTACT 2.0 analysis with an assigned host, 35 had a host predicted using WIsH, of which only 23 (66%) and 8 (23%) was the same as the homologue-based prediction at the order and family levels, respectively. These analyses therefore

756 indicated that while the *k*-mer based approach using WIsH was partially successful in identifying correct linkages at the family level, it was not robust for identifying linkages with a high level of taxonomic resolution or confidence.

758

Gene-sharing networks of viral metagenomic contigs

760 Predicted hosts of mVCs from this study (using CRISPR and homologue-based approaches) were generally inconsistent with the known hosts of RefSeq viruses that
762 were linked through vConTACT 2.0 gene-sharing network analysis⁸. However, mVCs from this study placed into individual clusters all had the same predicted host
764 (Supplementary Figure 4). Initial networks used mVCs with a conservative host prediction only (i.e. a minimum of ≥ 5 homologues linked to one taxonomic family).
766 With the exception of one cluster of 3 mVCs, which were predicted to be linked to *Methylococcaceae* and non-methane oxidising methylotrophic *Hyphomicrobiaceae*,
768 all individual networks were restricted to one family for methanotrophs or one cluster of exclusively rhizobia-linked mVCs. Further analyses included all mVCs >10 kbp in
770 this study (i.e. including those with <5 host associated homologues) and the same networks were identified i.e. mVCs linked within the same cluster all had homologues
772 (i.e. with <5 or ≥ 5) linking them to the same host family.

For *Methylocystaceae* mVCs, two separate but linked clusters were identified.
774 The first contained CRISPR-linked *Methylocystaceae*-associated mVCs, which were recovered from both pH 4.5 and 7.5 soil (although individual mVCs were restricted to
776 one soil pH). A second cluster was dominated by mVCs from pH 4.5 soil only, indicating that most active *Methylocystaceae* viruses belonged to one of two distinct
778 lineages. There was a clear difference in the prediction of category-2 ('likely') and category-3 ('possible') mVCs associated with these two clusters. While category-3
780 viruses are often excluded prior to analysis of soil viromes⁹, CRISPR analysis demonstrated that one-third of linked mVCs were of the lowest category of confidence.
782 It must be recognised that a proportion of the predicted category-3 mVCs in this study will not be derived from viruses, and clusters composed exclusively of category-3
784 mVCs without further validation (e.g. without CRISPR spacer linkages) must be interpreted with caution. However, all major clusters identified through gene-sharing
786 network analysis contained a mixture of category-2 and -3 mVCs indicating that they represented groupings of genuine virus-derived genomes.

788 An intriguing finding was the linkage of host family-specific viruses from two
different geographical regions (Scotland and Sweden) and contrasting soil types
790 (agricultural loamy-sand and permafrost peatland soils), from this study and that of
Emerson et al.⁹, respectively. In the latter study, 13 of 1,907 mVCs were predicted to
792 have a methanotroph host, of which 9 were linked specifically to the *Methyocystaceae*.
Intriguingly, 7 of these were also linked to a predicted *Methyocystaceae* mVC in our
794 study, with one additional mVC predicted to have a *Methylocapsa* host (i.e. belonging
to the *Beijerinckiaceae* which is another methylotrophic family of the Rhizobiales) but
796 also contained one predicted gene with a 'best hit' match to a *Methylocystis* genome
homologue. As soil pH is recognised as one of the dominant factors driving microbial
798 community structures in soil¹⁰, it is interesting to note that linked mVCs from both
studies were also from acidic soils only.

800

References

- 802 1. Nicol, G.W. & Prosser, J.I. Strategies to determine diversity, growth and activity of
ammonia oxidising archaea in soil. *Meth. Enzymol.* **496**, 3-34 (2011).
- 804 2. McLaren M.R., Willis, A.D. & Callahan, B.J. Consistent and correctable bias in
metagenomic sequencing experiments. *eLife* **8**, e46923 (2019).
- 806 3. Chistoserdova, L., Kalyuzhnaya, M.G. & Lidstrom, M.E. The expanding world of
methylotrophic metabolism. *Annu. Rev. Microbiol.* **63**, 477–499 (2009).
- 808 4. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal
from microbial genomic data. *PeerJ* **3**, e985 (2015).
- 810 5. Ren, J. et al. Identifying viruses from metagenomic data by deep learning. *Quant.*
Biol. **8**, 64-77 (2020).
- 812 6. Trubl, G., Hyman, P., Roux, S & Abedon, S.T. Coming-of-age characterization of
soil viruses: a user's guide to virus isolation, detection within metagenomes, and
814 viromics. *Soil Syst.* **4**, 23 (2020).
7. Galiez, C, Siebert, M., Enault, F., Vincent, J. & Söding, J. WisH: who is the host?
816 Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**,
3113-3114 (2017).
- 818 8. Jang, H.B. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes
is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632-639 (2019).
- 820 9. Emerson, J.B. et al. Host-linked soil viral ecology along a permafrost thaw
gradient. *Nat. Microbiol.* **3**, 870-880 (2018).

- 822 10. Bartram, A.K. et al. Exploring links between pH and bacterial community
composition in soils from the Craibstone Experimental Farm. *FEMS Microbiol.*
824 *Ecol.* **87**, 403–415 (2014).