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

Sungeun Lee[1], Ella T. Sieradzki[2], Alexa M. Nicolas[3], Robin L. Walker[4], Mary K. Firestone[2,5], Christina Hazard[1#] and Graeme W. Nicol[1#]

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

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

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

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

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

[#] Corresponding author

Abstract

The concentration of atmospheric methane continues to increase with microbial communities controlling soil-atmosphere fluxes. While there is substantial knowledge of the diversity and function of organisms regulating methane production and consumption, the frequency and impact of interactions with viruses on their activity in soil is unknown. Metagenomic sequencing of soil microbial communities has enabled identification of linkages between viruses and hosts. However, determining host-virus 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 of assimilated carbon from active hosts to virus. Using DNA stable-isotope probing combined with metagenomic analyses, we characterized methane-fueled microbial networks in acidic and neutral pH soils, specifically primary and secondary utilizers of carbon, together with the recent transfer of methane-derived carbon to viruses. Sixty-three percent of viral contigs from replicated soil incubations contained genes associated with known methanotrophic and methylotrophic bacteria. Genomic sequences from ^{13}C-enriched viruses were present in clustered regularly interspaced short palindromic repeats (CRISPR) arrays of multiple, closely-related *Methylocystis* populations, revealing differences in their history of viral interaction. Viruses infecting non-methanotrophic methylotrophs and heterotrophic predatory bacteria were also identified through the analysis of shared

homologous genes, demonstrating that carbon is transferred to a diverse range of viruses associated with methane-fueled microbial food networks.

Main text

Microorganisms play a central role in global carbon (C) biogeochemical cycling in soil 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., 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 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 prokaryotic ecology in soil remains comparatively unknown (Emerson, 2019). In particular, difficulties remain in identifying the frequency of active interactions between 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-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 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 communities has enabled identification of linkages between viruses and hosts involved in carbon cycling both through identifying CRISPR spacer sequences in viral genomes and the presence of viral genes encoding enzymes involved in complex carbon degradation (Emerson et al., 2018). However, determining virus-host associations *in situ* with these methods does not elucidate the frequency of viral infections, with linkages potentially associated with populations not active under current conditions, or even relic DNA (Carini et al., 2017).

Methanotrophs are a critically important group in soil systems, removing 5% of atmospheric methane (Curry, 2007) and controlling fluxes to the atmosphere from methanogenic activity in anoxic compartments (Le Mer and Roger, 2001). Aerobic methanotrophs use CH_4 for both carbon and energy requirements and key representatives in soil belong to the type I Gammaproteobacteria family *Methylococcaceae*, type II Alphaproteobacteria families *Methylocystaceae* and *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 (Zhao et al., 2020). In addition, a wide variety of non-methanotrophic methylotrophs utilise methanol produced and excreted by

68 methanotrophs, and together methanotrophic and other methylotrophic single carbon compound (C1)-utilising consortia assimilate methane-derived carbon in a variety of habitats
70 (Chistoserdova et al., 2010).

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

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

Reads from individual metagenomes were assembled before taxonomic assignment of
96 individual contigs. Reproducibly distinct communities were enriched in the two soils (Supplementary Fig. 2), with *Methylocystaceae*- and *Methylococcaceae*-assigned sequences
98 representing 25.5% vs 4.6% and 0.1% vs 23.4% of contigs in pH 4.5 and 7.5 soils, respectively. Eleven bacterial families each contained >1% of assembled contigs, nine of which include
100 known C1-utilizing taxa (*Beijerinckiaceae*, *Bradyrhizobiaceae*, *Burkholderiaceae*, *Comamonadaceae Gemmatimonadaceae*, *Hyphomicrobiaceae*, *Methylococcaceae*,

102 *Methylocystaceae* and *Methylophilaceae*). We resolved twenty-three medium and high-quality (Bowers et al., 2017) metagenome-assembled genomes (MAGs) (Supplementary Table 2),
104 including 12 methanotrophs. Specifically, 3 MAGs represented gammaproteobacterial type I methanotrophs (*Methylobacter*) and 9 MAGs represented alphaproteobacterial type II
106 methanotrophs (*Methylocystis*, *Methylosinus* or *Methylocapsa*). Secondary utilisers of methane-derived organic carbon were also identified with 9 MAGs associated with established
108 or putative non-methanotrophic methylotrophs, lacking methane oxidation machinery but capable of utilizing methanotroph-derived methanol (see Supplementary Text). These included
110 representatives of the *Gemmatimonadales*, *Hyphomicrobium*, *Herminiimonas* and *Rudaea*, the latter two, to our knowledge, not having been previously associated with methylotrophy but
112 possessed predicted methanol and formate dehydrogenases (Supplementary Table 2). Two MAGs represented strains of *Bdellovibrio* and *Myxococcus*, known predatory bacteria,
114 indicating that growing methanotrophic and methylotrophic populations were preyed upon (Pérez et al., 2018).

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

124 mVCs were linked to host bacteria using three different approaches: identifying incorporation of viral DNA into spacers of bacterial CRISPR arrays, similarity of homologous
126 genes possessed by both host and virus contigs, and *k*-mer similarity between potential host and virus contigs, the last approach being considered only partially successful (see
128 Supplementary Text). CRISPR arrays were identified in 3 of 23 MAGs, each associated with the genus *Methylocystis* or *Methylosinus* of the *Methylocystaceae* (Fig. 1). In the acidic soil,
130 complete CRISPR arrays of growing methanotrophs were associated with two *Methylocystis* MAGs (MAG identifiers 5 and 6) sharing 79.2% average nucleotide identity (ANI) and likely
132 representing different species (Jain et al., 2018). A further six complete CRISPR arrays were identified in unbinned bacterial contigs all possessing the same direct repeat (DR) sequence.
134 These eight arrays varied in size ranging from 9 to 114 DRs and contained a total of 432 spacers. Comparison of spacer incorporation between arrays revealed that these multiple

136 closely-related populations had different histories of viral interaction and subsequent spacer incorporation. Genome sequences from ^{13}C-enriched viral populations were represented by
138 seven mVCs and matched 29.5% of spacers. In addition, 7.9% of spacers possessed a one nucleotide mismatch, all of which represented a synonymous substitution, indicating that
140 variation was the result of mutation in viral genome increasing the ability evade CRISPR-CAS defense systems and/or genetic variation in closely related viral populations. Only three pairs
142 of spacers were identical, with each pair member located on a different array. Potential variation in virus host range was observed, with 3 and 2 mVCs linked to one or both
144 *Methylocystis* MAGs, respectively, for example.

Surprisingly, a large number of spacers in individual *Methylocystis* CRISPR arrays
146 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 from *Methylocystis*-
148 associated viruses, mVCs were examined for host-specific conserved protospacer-adjacent motif (PAM) sequences (Mojica et al., 2009). Consistent with the identification of genuine
150 protospacers, 138 of linked 146 spacers (i.e. all possessing ≤ 1 mismatch) had the conserved PAM sequence 'TTC' (target-centric orientation) (Leenay and Beisel, 2017) with all PAM
152 sequences indicating a consensus sequence of 'NTY'. The variation in spacer position between the arrays revealed temporal differences in virus infections. For example, the most recently
154 integrated spacer in 3 of 8 different *Methylocystis* CRISPRs was derived from a ^{13}C-enriched virus represented by mVC_12213_cat2, indicating that incorporation may have occurred
156 during the incubation of the experiment.

Further analysis of all complete CRISPR arrays (i.e. including those in unbinned
158 contigs) linked to ^{13}C-enriched viruses revealed that the majority of viruses were associated with methanotrophic populations (Fig. 2a). In total, 39 unique CRISPR arrays represented 11
160 different variants (i.e. each variant having a unique DR sequence) with 9 linked to *Methylocystaceae* or *Methylococcaceae* populations. Individual DR variants were restricted to
162 either pH 4.5 or 7.5 soil soil. Using 100% sequence identity in searches between CRISPR spacer and mVC protospacer sequences, 19 VirSorter-predicted mVCs were linked to CRISPR
164 arrays. 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
166 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
168 only higher confidence contigs may exclude a substantial proportion of *bona fide* methanotroph virus-derived contigs in uncharacterized environments such as soil.

170 Analysis of tetranucleotide frequencies (TETRA) (Wang et al., 2017) clustered the 21 mVCs into three groups that were associated with the *Methylocystaceae*, *Methylococcoceae* 172 and an unknown group (Fig. 2b). The majority of viruses infected members of the *Methylocystaceae* family, with those infecting populations of the *Methylocystis* and 174 *Methylosinus* genera restricted to acidic and neutral pH soils, respectively. TETRA correlation coefficients of all *Methylocystaceae*-linked viruses were in the same range both within and 176 between either genus, suggesting co-evolution with their host rather than genetic drift and divergence was the primary mechanism for defining specific associations with *Methylocystis* 178 or *Methylosinus* strains.

Identification of homologous genes in mVCs that were shared with prokaryotic 180 genomes were always consistent with host-virus linkages established using spacer sequences from MAG CRISPR arrays. Specifically, BLASTp searches of genes present in the 9 mVCs 182 linked to *Methylocystaceae* populations via CRISPR spacer sequences all contained 'best hits' (identity > 30%, e-value < 10^{-5}, bit score > 50 and query cover > 70%) to a minimum of 5 184 homologues also found in *Methylocystaceae* host genomes. This was therefore used as a criterion for establishing host-virus linkages. Sixty-three percent of mVCs contained a 186 homologue that was linked to genomes of known C1-utilizing bacteria, with 35% linked specifically to populations from the *Methylocystaceae*, *Methylococcaceae* or 188 *Hyphomicrobiaceae* (Fig. 3a). While analysis of bacterial homologues in mVCs identified the taxonomic family of the assumed dominant host, they also indicated that individual viruses 190 may infect hosts of other families of the same taxonomic order, including those at different trophic levels. Specifically, within the *Rhizobiales*, mVCs linked to *Methylocystaceae* also 192 contained homologues shared with methylotrophs (i.e. *Bradyrhizobiaceae*, *Methylobacteriaceae* and *Rhizobiaceae*) (Fig. 3b), indicating that viruses of methanotrophs 194 may also infect methylotrophs that are active at the same time.

CH_4-derived C was also transferred to viruses of secondary or tertiary utilizers. One 196 group of mVCs were linked to methylotrophic *Hyphomicrobiaceae* and a second to a phylogenetically diverse range of nitrogen-fixing Rhizobia i.e. *Bradyrhizobiaceae*, 198 *Phyllobacteriaceae* and *Rhizobiaceae*. These lineages contain known methylotrophs, methanol dehydrogenases have been identified in a range of rhizobial species (Huang et al., 2019) and 200 these mVCs also contained homologues found in the genomes of nodulating *Methylobacterium* strains (Green and Ardley, 2018). Viruses of predatory *Bdellovibrio* and *Myxococcales* bacteria 202 were predicted, consistent with the recovery of corresponding bacterial MAGs in ^{13}C-enriched DNA. One mVC (20210-cat_2) was linked to the genus *Bdellovibrio* (sharing 11 of 67 mVC

204 genes) and three category-3 mVCs (i.e. a possible virus) was linked to *Myxococcales* populations, containing gene homologous to four families within the order. The high isotopic
206 labelling of both heterotrophic predators (with no identifiable C1-utilizing capability) and their viruses indicate that the predators were feeding primarily on C1-utilisers, as carbon
208 incorporated by feeding on unlabelled bacteria would dilute the enrichment in predators. As such, it also indicates that predatory bacteria may have preference for preying upon growing
210 populations rather than the non-C1-utilzing majority.

Gene-sharing network analysis of mVCs with viruses in the NCBI RefSeq database and
212 other metagenome studies were analyzed using vConTACT 2.0 (Jang et al., 2019). Any linkages with RefSeq viruses typically had low scores (i.e. sharing a low number of
214 homologues) and were linked to viruses of hosts that were inconsistent with our homologue-based predictions (Supplementary Table 3). No linkages were observed with recently reported
216 giant viruses of methanotrophs in freshwater lakes (Chen et al., 2020). However, in a recent study of 197 metagenomes from Swedish peatland soil, Emerson et al. (2018) identified 13
218 viruses linked to methanotrophs. Intriguingly, 8 of these were linked in our viral gene-sharing network, with both studies predicting *Methylocystaceae* hosts using different methods of host
220 annotation (Supplementary Text, Supplementary Fig. 4) and revealing the distribution of specific *Methylocystaceae*-associated viral groups present in different geographical areas and
222 soil types. Analysis of gene-sharing networks of mVCs from this study indicated that there were two distinct *Methylocystaceae*-linked viral clusters with variation in both distribution in
224 soil and genome composition. Specifically, one cluster was associated with low pH only whereas the second cluster contained viruses found in both pH 4.5 and 7.5 soils, including
226 those linked by CRISPR spacer sequences. Individual networks of *Methylococcaceae*- and rhizobia-associated mVCs were also identified, typically associated with one of the two soils
228 of contrasting pH. Taxonomically-linked mVCs with ≥5 homologues were consistently placed in networks with other mVCs containing 1-4 homologues to the same methanotroph or
230 methylotroph families, confirming host linkage to a larger number of mVCs.

mVCs contained 8,174 genes, with 49.6% (4,054) annotated representing 606 unique
232 functions. Of these, genes encoding viral proteins accounted for 9.8% (397 genes) and included major capsid proteins, tail proteins, integrases, portal proteins and terminases. Bacterial
234 proteins used for viral replication accounted for 5.1% (206 genes). A number of metagenomic studies have demonstrated that viruses can possess genes encoding sub-unit C of ammonia or
236 particulate methane monooxygenases as auxiliary metabolic genes (AMGs) (Chen et al., 2020;

Ahlgren et al., 2019) which are also typically found as isolated genes in genomes in addition
238 to being present in clusters or operons encoding A and B sub-units (Nicol and Schleper, 2006).
In this study, one low confidence mVC (7.3 kbp, category-3) was identified as containing an
240 isolated *pmoC* gene that was phylogenetically related to growing *Methylocystis* populations
but was distinct from *pmoC* sequences found in viruses associated with sediment *Methylocystis*
242 populations (Supplementary Fig. 5).

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

252 **Methods**

Soil microcosms

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

266 *DNA-SIP*

Genomic DNA was extracted from 0.5 g soil samples using a CTAB buffer phenol:chloroform:
268 isoamyl alcohol bead-beating protocol and subjected to isopycnic centrifugation in CsCl
gradients, recovery and purification as previously described (Nicol and Prosser, 2011). Briefly,
270 6 ug of genomic DNA was added to 8 ml CsCl-Tris EDTA solution (refractive index (RI) of

1.4010; buoyant density of 1.71 g ml^{-1}) in polyallomer tubes before sealing and
272 ultracentrifugation at 152,000 × *g* (50,000 rpm) in a MLN80 rotor (Beckman-Coulter) for 72 h
at 25°C. CsCl gradients were fractionated into 350 ul aliquots using an in-house semi-
274 automated fraction recovery system before determining RI and recovering DNA. The relative
abundance of bacterial 16S rRNA and methanotrophic *pmoA* genes in genomic DNA
276 distributed across the CsCl gradients was determined by qPCR in a Corbett Rotor-Gene 6000
thermocycler (Qiagen) using primer sets P1(341f)/P2(534r) (Muyzer et al., 1993) and
278 A189F/A682R (Holmes et al., 1995) respectively. Twenty-five µl reactions contained 12.5 µl
2X QuantiFast SYBR Green Mix (Qiagen), 1 µM of each primer, 100 ng of T4 gene protein
280 32 (Thermo Fisher), 2 µl 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 of 15 min at
282 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 for the *pmoA* gene assay,
284 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 buoyant density >1.732 g ml^{-1} were then
286 pooled for each ^{12}C- and ^{13}C-CH_4-incubated replicate for 16S rRNA gene amplicon sequence
and metagenomic analysis.
288

Metagenome sequencing, assembly & annotation

290 Library preparation and sequencing was performed at the Joint genome Institute (JGI),
Berkeley, USA. Libraries were produced from fragmented DNA using KAPA Biosystems
292 Library Preparation Kits (Roche) and quantified using KAPA Biosystems NGS library qPCR
kits. Indexed samples were sequenced (2 x 150 bp) on the Illumina NovaSeq platform with
294 NovaSeq XP v1 reagent kits and a S4 flowcell. Raw reads were processed with JGI's
RQCFilter2 pipeline that utilised BBTools v38.51 (Bushnell, 2016). Reads containing adapter
296 sequences were trimmed and those with ≥ 3 N bases or ≤ 51 bp or ≤ 33% of full-read length
were removed along with PhiX sequences using BBDuk, and reads mapped to human, cat, dog
298 or mouse references at 95% identity were removed using BBMap. *De novo* contig assembly of
the 100 - 196 million quality-controlled reads per metagenome was performed using
300 MetaSPAdes v3.13.0 (Nurk et al., 2016). The 1 - 2 million contigs per metagenome were then
concatenated together, and contigs larger than 5 kbp were dereplicated at 99% average
302 nucleotide identity (ANI) using PSI-CD-HIT v4.6.1 (Fu et al., 2012) and binned using
MetaWRAP v1.2.1 (Uritskiy et al., 2018) (Supplementary Table 4). Bin completion and
304 contamination was determined by CheckM v1.0.12 (Parks et al., 2015). Taxonomic annotation

of contigs was performed using Kaiju (Menzel et al., 2016) with the NCBI RefSeq database
306 (Release 94; 25 June 2019) (O'Leary et al., 2016) and MAGs using GTDB-Tk v0.3.2
(Chaumeil et al., 2019) with the Genome Taxonomy Database (release 89, 21 June 2019) (Parks
308 et al., 2018). Protein sequence annotation was performed using InterProScan 5 (e-value < 10^{-5}) (Jones et al., 2015). Taxon-annotated GC mol% coverage plots for each of the six
310 metagenomes were produced using the MetaWRAP-Bloblogy module (Uritskiy et al. 2018).
Pairwise ANI comparison of MAGs was calculated using FastANI (Jain et al., 2019).
312

Amplicon sequencing and analysis

314 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 previously described
316 (Finn et al., 2020). Reads with a quality score < 20 and length < 100 bp were discarded using
FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). High-quality reads were
318 merged using PANDAseq v2.11 (Masella et al., 2012), and denoising and chimera removal
performed with UNOISE3 (Edgar, 2016). Amplicon sequence variants (ASVs) were annotated
320 using the RDP classifier v2.11 (Wang et al, 2007). Non-metric multidimensional scaling of
Bray-Curtis dissimilarity derived from the relative abundance of ASVs was performed with the
322 metaMDS function in the vegan package (Oksanen et al., 2019) in R v3.6.0.

324 *Virus prediction*

Metagenomic viral contigs (mVCs) were predicted from 9,190 contigs > 10 kbp using
326 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) was also used to
328 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 (Supplementary Table 5). The
330 relative abundance of each mVC in the six metagenomes was determined using the
MetaWRAP-Quant_bins module (Uritskiy et al., 2018) and a heatmap produced using the
332 heatmaply package in R.

334 *Virus-host linkage*

CRISPR arrays within MAGs and unbinned contigs were identified using the CRISPR
336 Recognition Tool v1.2 (Bland et al., 2007) (Supplementary Table 6). DR and spacer sequences
were extracted before performing 100% identity searches against positive and negative strands
338 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 spacer sequences in mVCs, 10
340 nucleotides before and after the spacer sequence were extracted to identify associated host-specific PAM sequences. Conserved and variant PAM sequences were manually identified.
342 Correlation coefficients of pairwise comparison of the tetra-nucleotide frequencies (TETRA) between unique CRISPR-associated mVCs were calculated using Python package pyani
344 v0.2.10 (Pritchard et al., 2016). To identify homologous genes shared between CRISPR-linked viruses and hosts, gene prediction was performed using Prodigal v2.6.3 (Hyatt et al., 2010)
346 with the -p meta option followed by protein alignment with Blastp (identity > 30%, e-value < 10^{-5}, bit score > 50 and query cover > 70%) and protein sequence annotation using InterProScan
348 5 (e-value < 10^{-5}). Gene homology between all mVCs and prokaryotes in the NCBI nr database was determined using Diamond Blastp (e-value < 10^{-5}) (Buchfink et al., 2015). Virus-host
350 prediction using k-mer frequencies was performed with WIsH v1.0 (p-values < 0.05) (Galiez et al., 2017). Networks based on shared gene content was constructed using vConTACT 2.0
352 (Jang et al., 2019) with the NCBI RefSeq database (Release 94; 25 June 2019).

354 *Phylogenetic analysis of* PmoC *and* PxmC *protein sequences*

Maximum likelihood analysis of inferred protein sequences of membrane-bound
356 monooxygenase C sub-units from methanotroph MAGs and reference sequences (Supplementary Table 7) was performed on unambiguously aligned sequences using PhyML
358 3.0 (Guindon et al., 2010) with automatic model selection (LG substitution, gamma distribution (0.06) and proportion of invariable sites (0.087) estimated). Bootstrap support was calculated
360 from 100 replicates.

362 **Data availability**

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

368 **Acknowledgments**

The sequencing data were generated under JGI Community Science Program proposal 503702
370 awarded to GWN and CH. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science
372 of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This work was

funded by an AXA Research Chair awarded to GWN and a France-Berkeley Fund grant (18-2019) awarded to GWN and MKF. The pH gradient experiment is funded through the Scottish Government RESAS 2016-21 programme. The authors would like to thank Dr. Joanne Emerson for valuable discussion.

Author contributions

The research program was conceived by and funded from grants awarded to GWN, 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 approaches, discussed data and commented on the manuscript. RW coordinated soil sampling and commented on the manuscript. All authors approved the manuscript.

386 **References**

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

390 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 (2012).

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

394 Bowers, R.M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotech.* **35**, 725-731
396 (2017).

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

Bushnell, B. BBTools software package. http://sourceforge net/projects/bbmap (2016).
400

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

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

406 Chen, L.-X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat. Microbiol.* In press (2020).
408

Chistoserdova, L., Kalyuzhnaya, M.G. & Lidstrom, M.E. The Expanding World of
410 Methylotrophic Metabolism. *Annu Rev Microbiol.* **63**, 477–499 (2009).

412 Curry, C.L. Modeling the soil consumption of atmospheric methane at the global scale, *Global Biogeochem. Cycles* **2**1, GB4012 (2007).

414

Edgar, R.C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
416 sequencing. Preprint at https://doi.org/10.1101/081257 (2016).

418 Emerson, J.B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol*. **3**, 870-880 (2018).

420

Emerson, J.B. Soil Viruses: A New Hope *mSystems* **4**, e00120-19 (2019).

422

Finn, D.R., Lee, S., Lazén, M.B., Nicol, G.W. & Hazard, C. Cropping systems that improve
424 richness convey greater resistance and resilience to soil fungal, relative to prokaryote, communities. Preprint at https://doi.org/10.1101/2020.03.15.992560 (2020).

426

Frossard, A., Hammes, F., & Gessner, M.O. Flow cytometric assessment of bacterial
428 abundance in soils, sediments and sludge. *Front. Microbiol.* **7**, 903 (2016).

430 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

432 Galiez, C, Siebert, M., Enault, F., Vincent, J. & Söding, J. WisH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113-3114 (2017).

434 Gómez, P. & Buckling. Bacteria-Phage Antagonistic Coevolution in Soil. *Science* **332**, 106-109 (2011).

436 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 a new genus,
438 *Methylorubrum* gen. nov. *Int. J. Syst. Evol. Microbiol.* **68**, 2727-2748 (2018).

Guindon, S. et al. New Algorithms and Methods to Estimate Maximum-Likelihood
440 Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol*. **59**, 307-21 (2010).

442 Holmes, A.J., Costello, A., Lidstrom, M.E. & Murrell, J.C. Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS*
444 *Microbiol Lett.* **132**, 203-208 (1995).

446 Huang, J. et al. Rare earth element alcohol dehydrogenases widely occur among globally distributed, numerically abundant and environmentally important microbes. *ISME J.* **13**, 2005-
448 2017 (2019).

Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site
450 identification. *BMC Bioinform*. **11**, 119 (2010).

Ignacio-Espinoza, J. C., Ahlgren, N.A. & Fuhrman, J.A. Long-term stability and Red Queen-
452 like strain dynamics in marine viruses. *Nat. Microbiol*. **5**, 265-271 (2020).

Jain, C., Rodriguez-R, L.M., Phillippy, A.M. & Konstantinidis, K.T. High throughput ANI
454 analysis of 90K prokaryotic genomes reveals clear species boundarie. *Nature Comm.* **9**, 5114 (2018).

456 Jang, H.B. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol*. **37**, 632-639 (2019).

458 Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2015).

460 Kemp, J.S., Paterson, E., Gammack, S.M., Cresser, M.S. & Killham, K. Leaching of genetically modified *Pseudomonas fluorescens* through organic soils: influence of
462 temperature, soil pH, and roots. *Biol. Fert. Soils* **13**, 218–224 (1992).

464 Kirschke, S., et al. Three decades of global methane sources and sinks. *Nat. Geosci.* 6, 813–823 (2013).
466

Knief, C. Diversity and habitat preferences of cultivated and uncultivated aer- obic
468 methanotrophic bacteria evaluated based on *pmoA* as molecular marker. *Front. Microb.* **6**, 1346 (2015).
470

Le Mer, J. & Roger, P., Production, oxidation, emission and consumption of methane by soils:
472 a review. *Eur. J. Soil Biol.* **37**, 25–50 (2001).

474 Leenay, R.T. & Beisel, C.L. Deciphering, Communicating, and Engineering the CRISPR PAM. J. Mol. Biol. **429,** 177-191 (2017).
476

Lyu, Z., Shao, N., Akinyemi, T. & Whitman, W.B. Methanogenesis. *Curr. Biol.* **28**, 727–732
478 (2018).

Masella, A.P., Bartram, A.k., Truszkowski, J.M., Brown, D.G. & Neufeld, J.D. PANDAseq:
480 paired-end assembler for illumina sequences. *BMC Bioinform.* **13**, 31 (2012).

Menzel, P., Ng, K.L. & Krogh, A. Fast and sensitive taxonomic classification for
482 metagenomics with Kaiju. *Nat. Comm*. **7**, 11257 (2016).

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. Short motif
484 sequences determine the targets of the prokaryotic CRISPR defence system. *Microb.* **155,** 733-740 (2009).

486 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
488 genes coding for 16S rRNA. *Appl. Environ. Microbiol*. **59**, 695-700 (1993).

490 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).
492

Nicol, G.W. & Schleper C. Ammonia-oxidising Crenarchaeota: important players in the
494 nitrogen cycle? *Trends Microbiol*. **14**, 207–212 (2006).

496 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P.A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. **27**, 824-834 (2017).

498 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).

500 Oksanen, J. et al. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan (2019).
502

Paez-Espino, D., Pavlopoulos, G.A., Ivanova, N.N. & Kyrpides, N.C. Nontargeted virus
504 sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Proto*c. **12**, 1673
(2017).
506

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
508 assessing the quality of microbial genomes recovered from isolates, single cells, and
metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

510 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).

512 Pérez, J. Moraleda-Muñoz, A., Marcos-Torres, F.J. & Muñoz-Dorado, J. Bacterial predation:
75 years and counting! *Environ. Microbiol.* **18**, 766–779 (2018).

514 Pratscher, J., Vollmers, J., Wiegand, S., Dumont, M.G., & Kaster, A.-K. Unravelling the
identity, metabolic potential and global biogeography of the atmospheric methane-oxidizing
516 Upland Soil Cluster α. *Environ. Microbiol.* 20, 1016-1029 (2018).

Pasulka, A.L. et al. Interrogating marine virus-host interactions and elemental transfer with
518 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
520 taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal.
Methods 8, 12-24 (2016).
522

Radajewski, S., et al. Identification of active methylotroph populations in an acidic forest soil
524 by stable-isotope probing. *Microbiol*. 148, 2331–2342 (2002).

526 Ren, J. et al. Identifying viruses from metagenomic data by deep learning. *Quant. Biol.* **8,** 64-
77 (2020).
528

Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
530 microbial genomic data. *PeerJ* **3**, e985 (2015).

532 Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).

534

Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q
536 File Manipulation. *PLoS ONE* **11**, e0163962 (2016).

538 Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).

540

Uritskiy, G.V., DiRuggiero, J. & Taylor, J. MetaWRAP- a flexible pipeline for genome-
542 resolved metagenomic data analysis. *Microbiome* **15**, 158 (2018).

544 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* **1**,
546 e00009-15 (2015).

548 Williamson, K.E., Fuhrmann, J.J., Wommack, K.E., & Radosevich, M. Viruses in Soil Ecosystems: An Unknown Quantity Within an Unexplored Territory. *Annu. Rev. Virol.* **4**, 201-
550 219 (2017).

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

554

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naïve Bayesian Classifier for Rapid
556 Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl. Environ. Microbiol.* **73**, 5261-5267 (2007).

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

560

562 **Figure legends**

564 **Fig. 1.** Linkage of active ^{13}C-enriched viruses to *Methylocystaceae* populations in soil by comparison of spacer sequences in complete CRISPR arrays. a) Distribution of spacers from 7
566 mVCs in *Methylocystis* CRISPR arrays (MAGs 5, 6 and six unbinned contigs). b) CRISPR array of *Methylosinus* MAG 11 containing spacers linked to two mVCs. CRISPR array names
568 describe the individual soil microcosm that the contig was recovered from. Direct repeats are numbered (in grey), with the spacer after direct repeat 1 being the most recently incorporated.
570 Spacers with 100% identity or 1 mismatch to sequences in mVCs are represented by colour-coded squares and circles, respectively. Striped boxes representing sequences found in two
572 different mVCs.

574 **Fig. 2.** Presence and linkages of ^{13}C-enriched viruses to methanotroph CRISPR arrays in pH 4.5 and 7.5 soil. a) Presence of spacers from 21 mVCs in 11 different CRISPR array variants
576 (unique DR sequence). Taxonomic affiliation of CRISPR arrays to host families was determined by phylogenomic analysis of affiliated MAGs (1,6) or unbinned contigs (2-5, 6-9),
578 or inferred from shared homologues between linked mVCs and bacterial genomes (10, 11). All mVCs were >10 kbp except mVC_08964_cat.3 (9.8 kbp) and mVC_28139_DVF (5.1 kbp).
580 These two mVCs were also the only two predicted using DeepVirFinder, with calculated probabilities describing 'likely' and 'probable' viruses, respectively. b) TETRA correlation
582 coefficients between 21 CRISPR-linked mVCs. Colored-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
586 bacterial host populations through identification of shared homologous genes. a) Association of viruses with different bacterial families and functional groups inferred from the presence
588 of ≥ 5 shared homologous genes in category-1, -2 and -3 VirSorter-predicted mVCs. b) Proportion of homologues in methanotroph, methylotroph or predator viruses linked to other
590 bacterial families. Each chart summarizes those mVCs that all contain ≥ 5 homologues to one family (number of mVCs given in parentheses) but with other taxonomic linkages also given.
592 'Other' describes the proportion found in families each represented by less than <5% of homologues or those not annotated to the family level.

594 .

596 **Supplementary Fig. 1.** Buoyant density distribution of genomic DNA from total bacterial 16S rRNA genes and methanotroph communities possessing particulate methane monooxygenase 598 sub-unit A (*pmoA*) genes after isopycnic centrifugation in CsCl gradients. Genomic DNA was extracted from triplicate pH 4.5 and 7.5 soil microcosms incubated with a 10% ^{12}C- or ^{13}C- 600 CH_4 headspace. Vertical error bars are the standard error of the mean relative abundance and horizontal bars (mostly smaller than the symbol size) the standard error of the mean buoyant 602 density of individual fractions derived from three independent CsCl gradients, each representing an individual microcosm.

604

Supplementary Fig. 2. Taxonomic affiliation of contigs derived from ^{13}C-enriched high 606 buoyant density DNA recovered from methane-fuelled prokaryotic communities in triplicate pH 4.5 and pH 7.5 soil microcosms. Individual contigs are plotted as abundance vs GC mol%. 608 a) Taxonomic affiliation of contigs from families each representing >1% relative abundance. b) Taxonomic affiliation of 23 high- and medium-quality MAGs, with family or class level 610 given if the genus was not identified. Values in parenthesis beside each bin describe estimated genome completeness and contamination.

612

Supplementary Fig. 3. Heat-map of the relative abundance of 270 mVCs >10 kbp in length 614 from ^{13}C-enriched viral DNA derived from triplicate pH 4.5 and pH 7.5 soil microcosms. Normalized abundance is presented as reads per kbp.

616

Supplementary Fig. 4. Gene sharing network analysis of mVCs (one representative per 618 vOTU) from ^{13}C-enriched viruses in pH 4.5 and 7.5 soils. a) Taxonomic affiliation of hosts was predicted by homologue analysis, with mVCs containing ≥ 5 or < 5 linked homologues 620 highlighted. Eight mVCs from peatland soils linked to the *Methylocystaceae* (Emerson *et al.*, 2018) are also shown. b) Distribution of mVCs in pH 4.5 and 7.5 soil determined by the mean 622 ratio of normalized relative abundance from triplicate samples. Seven mVCs from peatland were only found in soils with pH ≤ 4.7. c) VirSorter category prediction and linkage to CRISPR 624 arrays via spacer sequence analysis.

626 **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 628 and one potential viral-derived contig. MAG-derived sequences are described by MAG number, genus and contig identifier (see Supplementary Table 7). Sequences in reference

630 methanotroph genomes and sediment-derived viruses (Chen et al., 2020) were included with NCBI accession numbers given in parenthesis. Circles at nodes describe percentage bootstrap
632 support from 100 replicates and the scale bar represents 0.02 changes per amino acid position.

634

636

Fig. 1

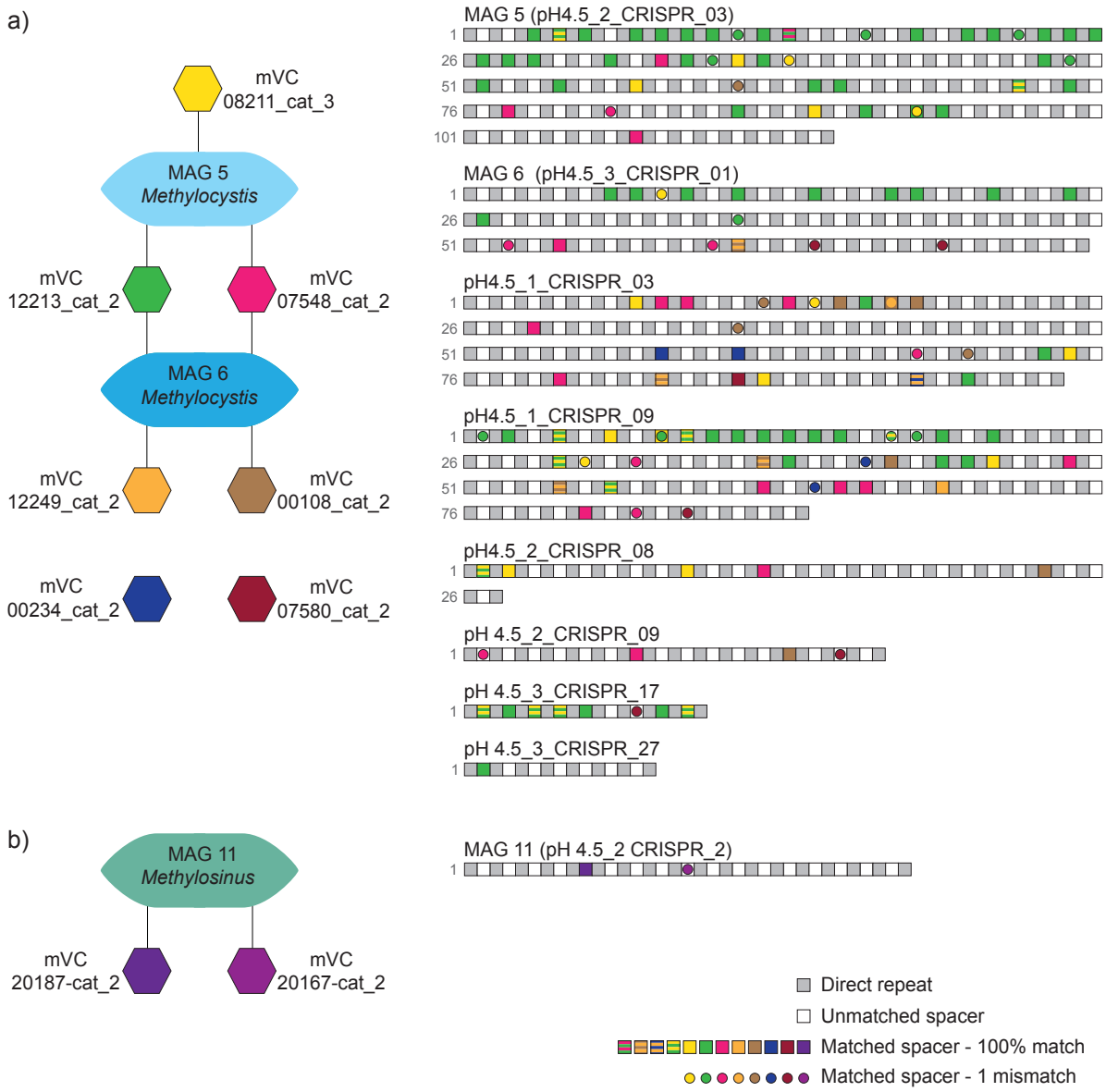

a)
mVC
08211_cat_3
MAG 5
Methylocystis
mVC
12213_cat_2
mVC
07548_cat_2
MAG 6
Methylocystis
mVC
12249_cat_2
mVC
00108_cat_2
mVC
00234_cat_2
mVC
07580_cat_2
MAG 5 (pH4.5_2_CRISPR_03)
MAG 6 (pH4.5_3_CRISPR_01)
pH4.5_1_CRISPR_03
pH4.5_1_CRISPR_09
pH4.5_2_CRISPR_08
pH 4.5_2_CRISPR_09
pH 4.5_3_CRISPR_17
pH 4.5_3_CRISPR_27
b)
MAG 11
Methylosinus
mVC
20187-cat_2
mVC
20167-cat_2
MAG 11 (pH 4.5_2 CRISPR_2)
Direct repeat
Unmatched spacer
Matched spacer - 100% match
Matched spacer - 1 mismatch

Fig. 2

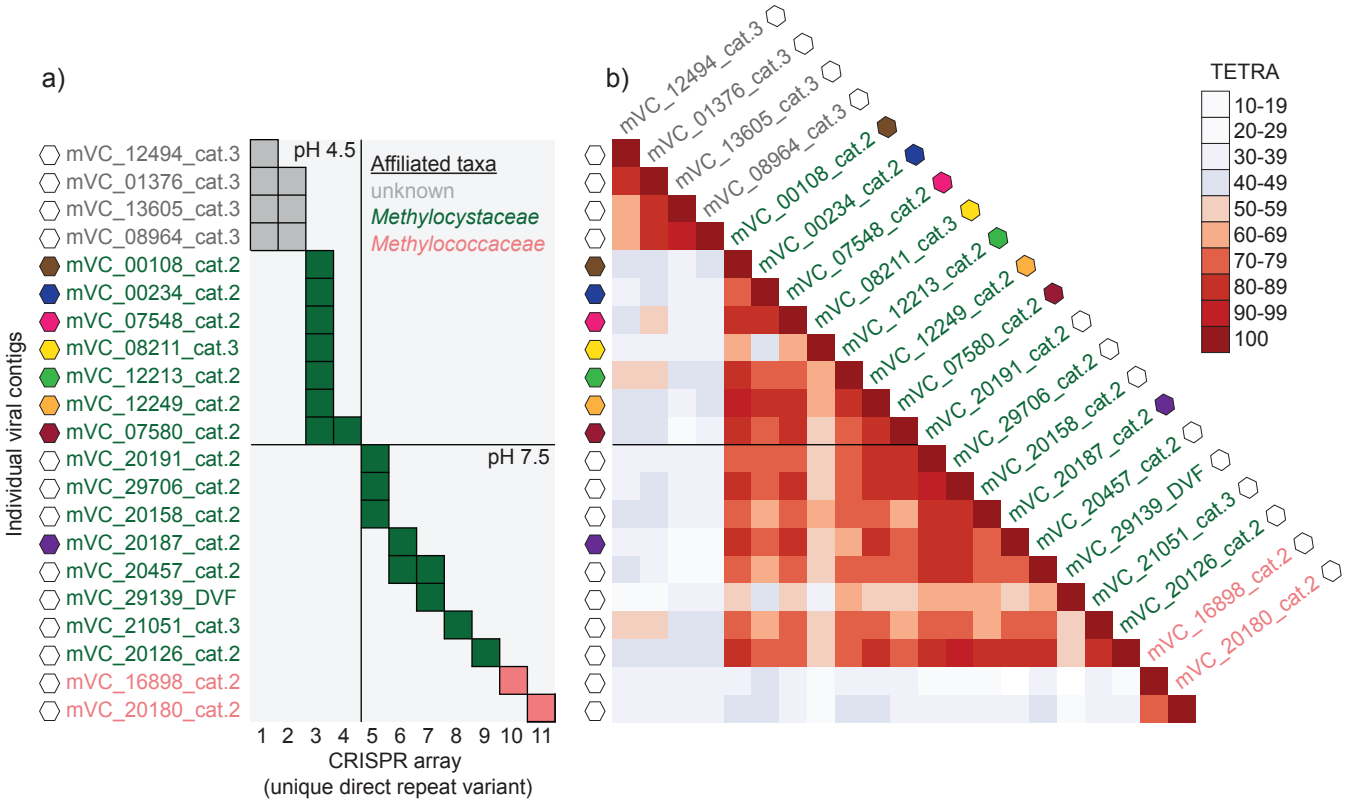

a)
b)
Individual viral contigs
mVC_12494_cat.3
mVC_01376_cat.3
mVC_13605_cat.3
mVC_08964_cat.3
mVC_00108_cat.2
mVC_00234_cat.2
mVC_07548_cat.2
mVC_08211_cat.3
mVC_12213_cat.2
mVC_12249_cat.2
mVC_07580_cat.2
mVC_20191_cat.2
mVC_29706_cat.2
mVC_20158_cat.2
mVC_20187_cat.2
mVC_20457_cat.2
mVC_29139_DVF
mVC_21051_cat.3
mVC_20126_cat.2
mVC_16898_cat.2
mVC_20180_cat.2
pH 4.5
pH 7.5
Affiliated taxa
unknown
Methylocystaceae
Methylococcaceae
1 2 3 4 5 6 7 8 9 10 11
CRISPR array
(unique direct repeat variant)
TETRA
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99
100

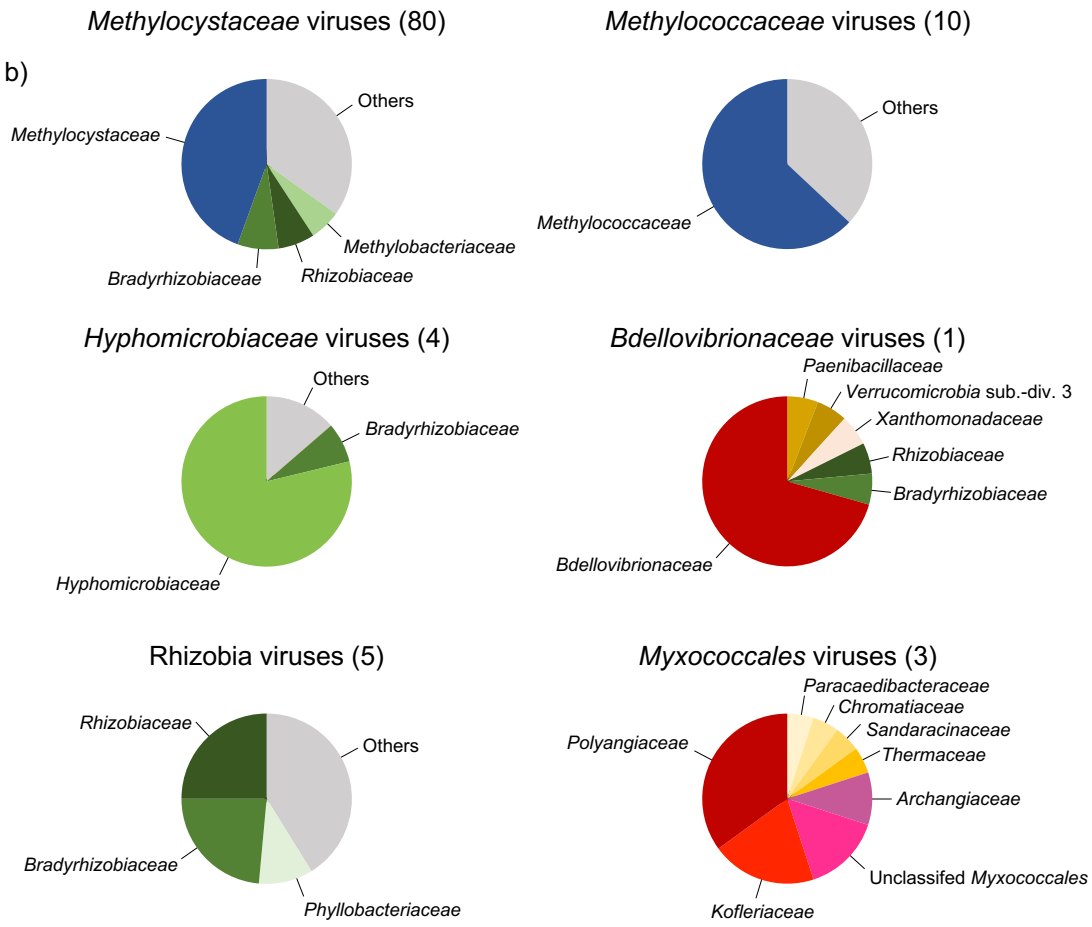
Fig. 3

a)

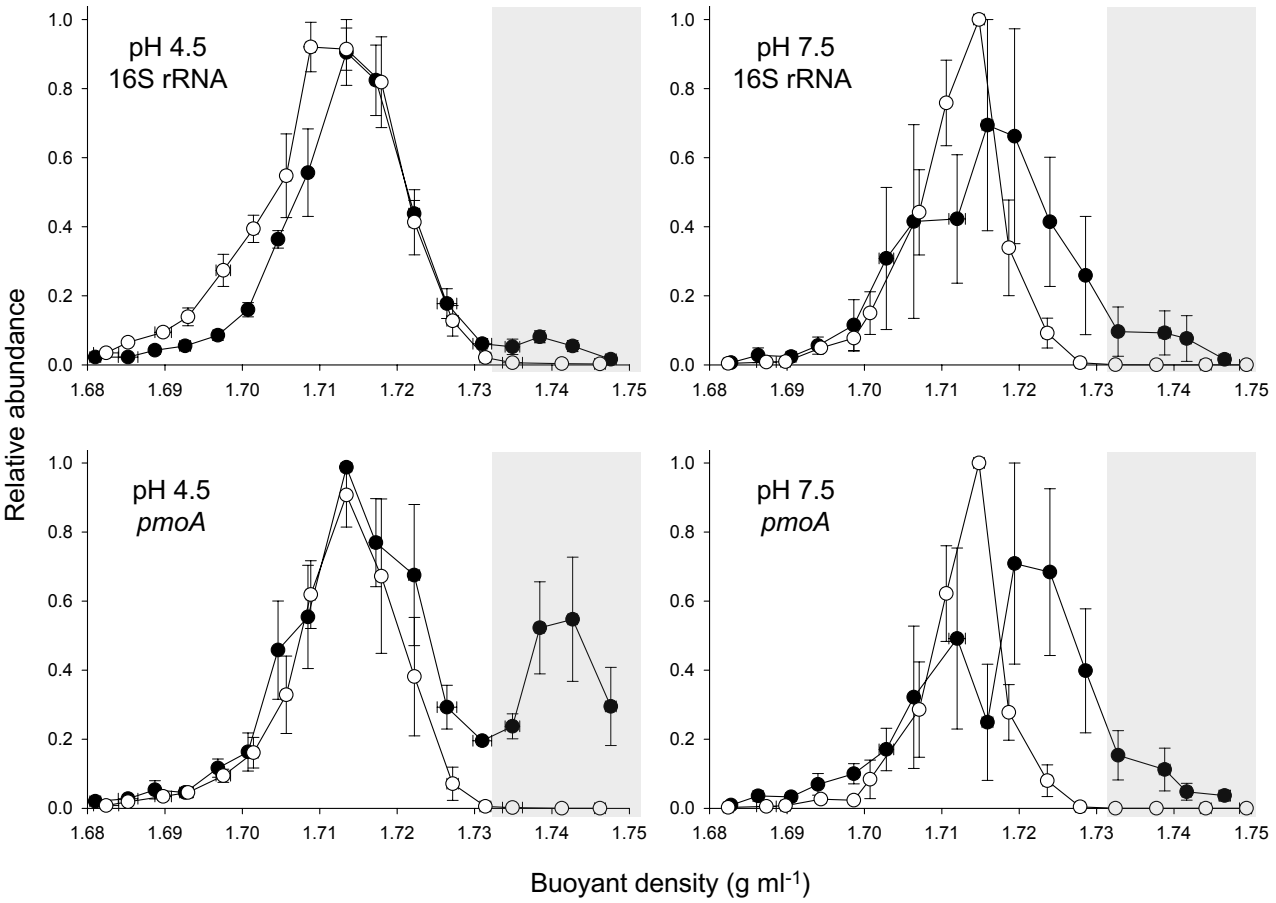
Order	Family	Cat 1	Cat 2	Cat 3
Rhizobiales	*Methylocystaceae*	●	●	●
	Hyphomicrobiaceae	○	●	●
	Various Rhizobia families	●	●	○
Methylococcales	*Methylococcaceae*	○	●	●
Myxococcales	Various families	○	○	●
Bdellovibrionales	*Bdellovibrionaceae*	○	●	○

● Methanotroph viruses
● Methylotroph viruses
● Predator viruses

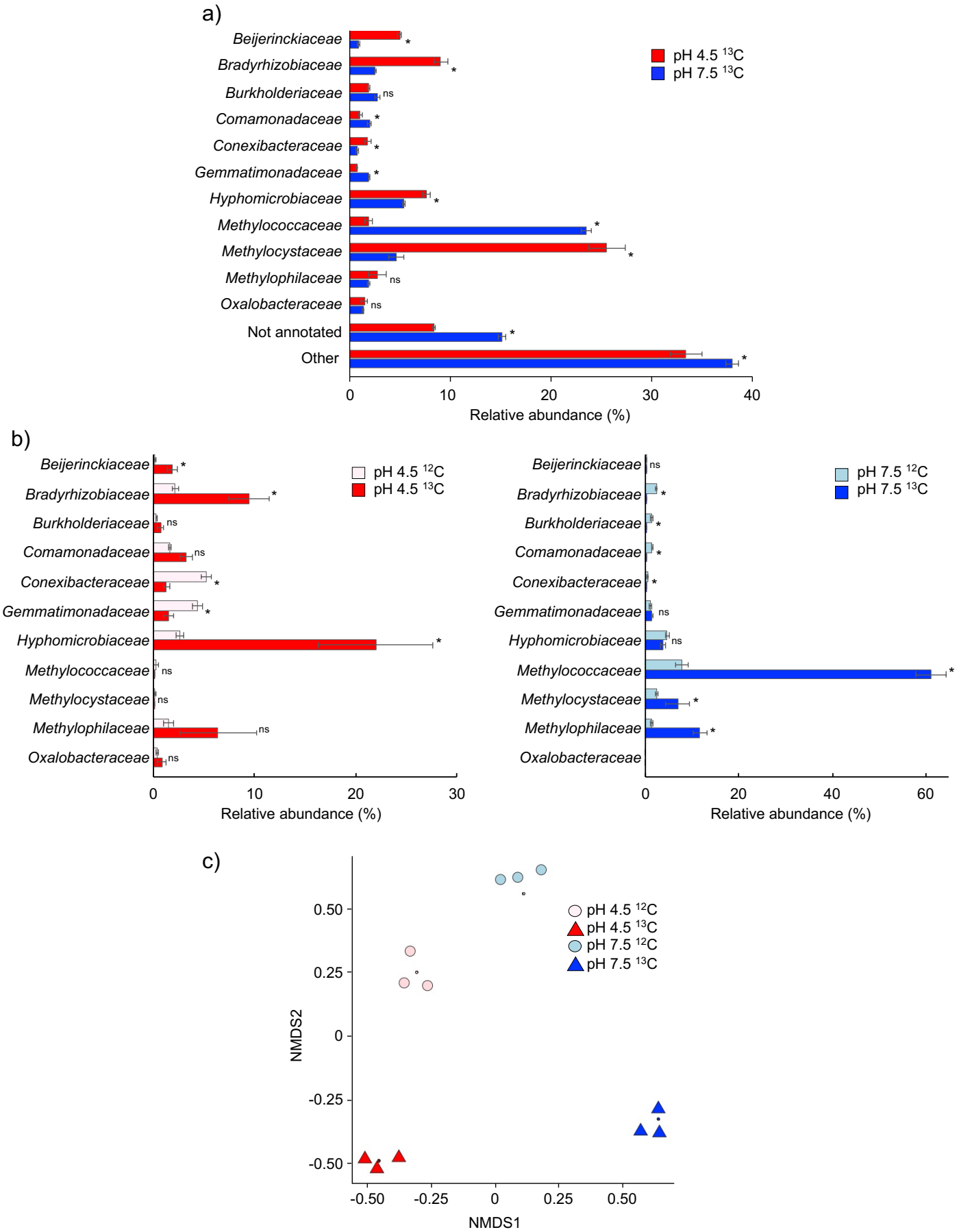
b)



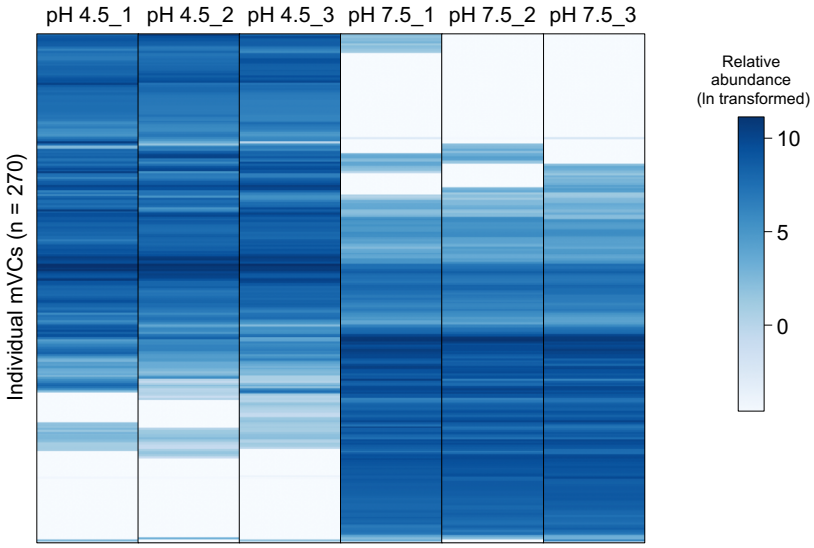
Supplementary Fig. 1



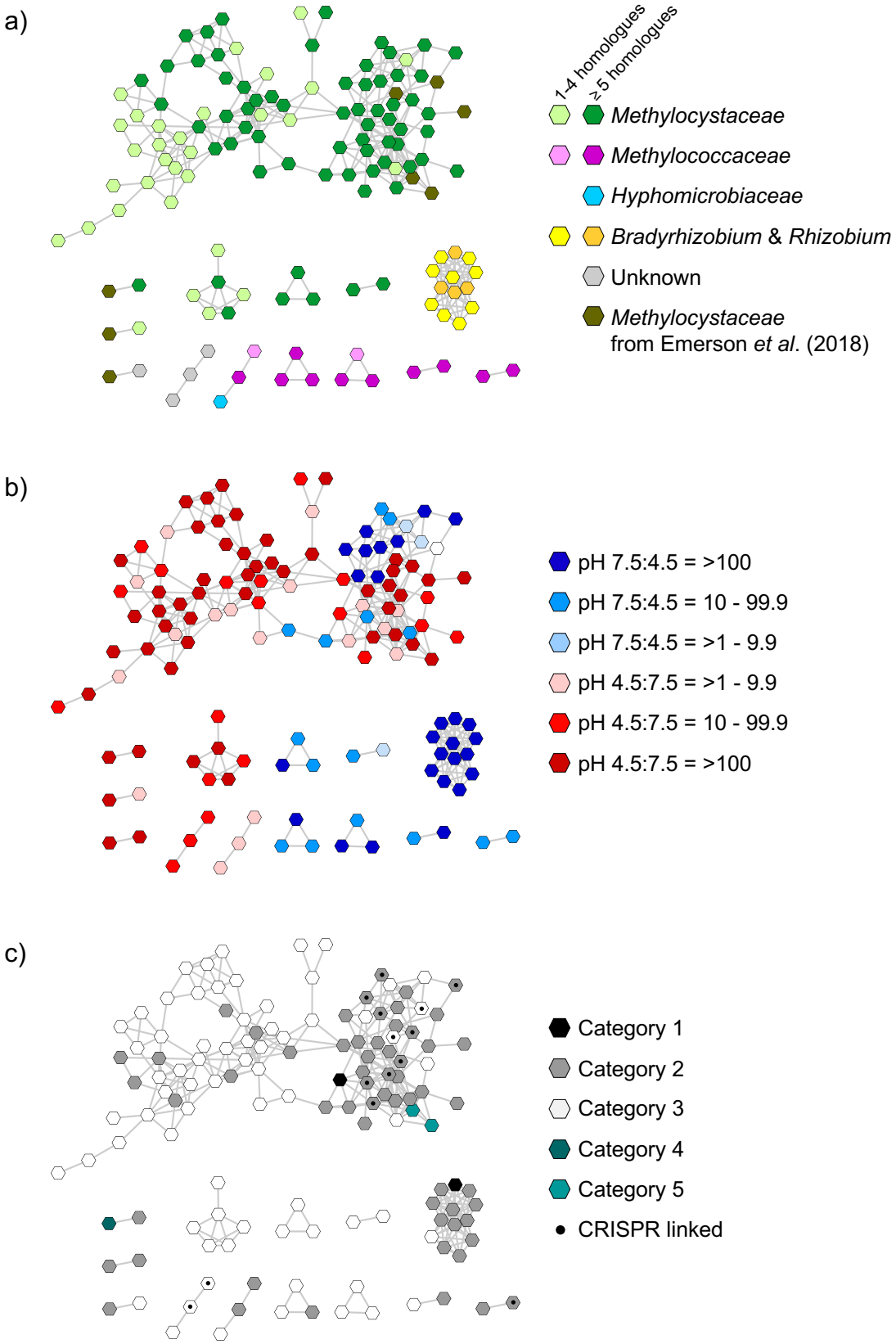
Supplementary Fig. 2



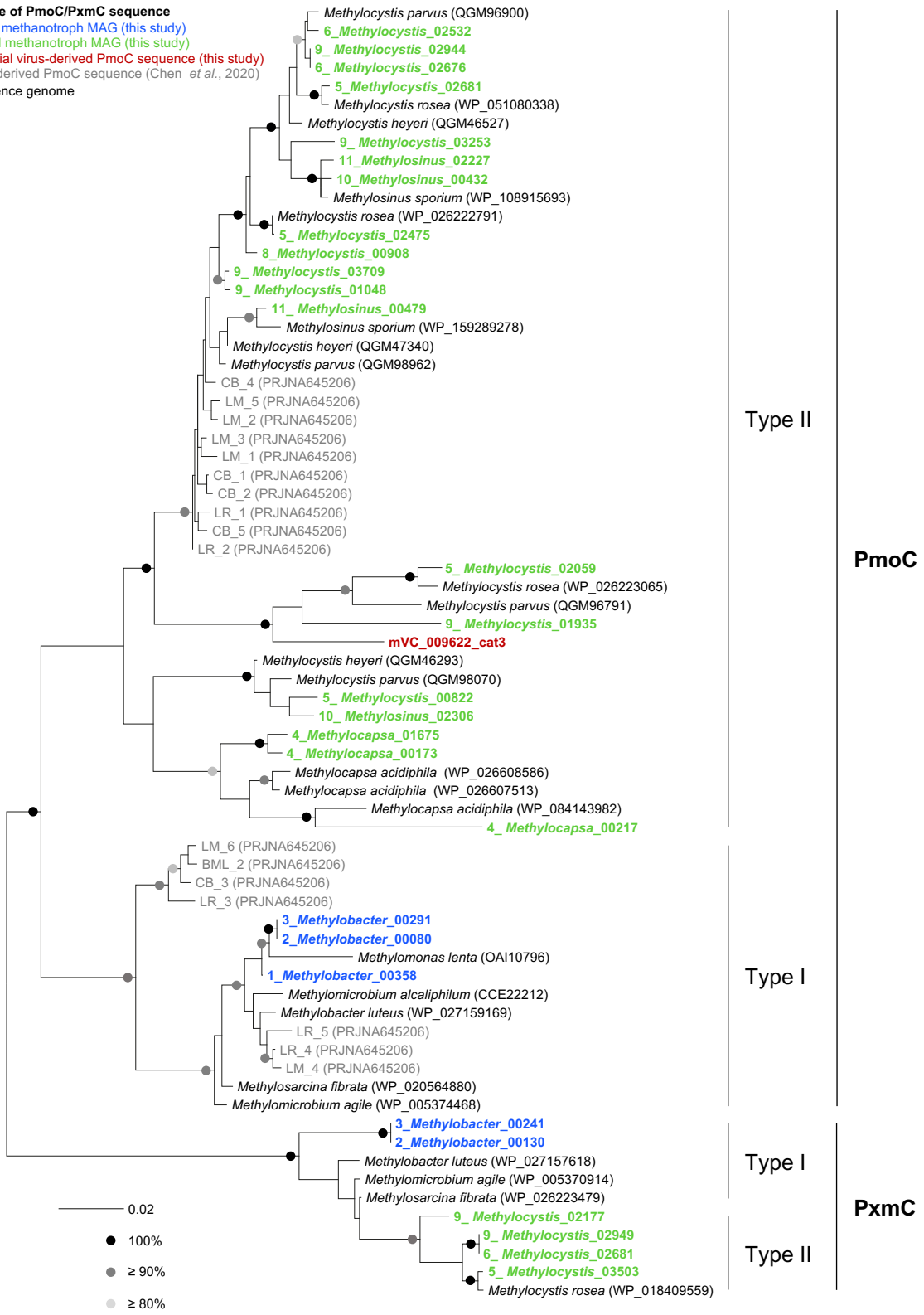
Supplementary Fig. 3



Supplementary Fig. 4



Supplementary Fig. 5



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

Sungeun Lee[1], Ella T. Sieradzki[2], Alexa M. Nicolas[3], Robin L. Walker[4], Mary K. Firestone[2,5], Christina Hazard[1#] and Graeme W. Nicol[1#]

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

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

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

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

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

corresponding authors

Email: graeme.nicol@ec-lyon.fr

Email: christina.hazard@ec-lyon.fr

Supplementary text

Comparison of 16S rRNA gene amplicon and metagenomic libraries of high buoyant density DNA from ^{12}C and ^{13}C-CH_4 incubations

The relative abundance of annotated contigs belonging to different families in replicate metagenomic libraries was reproducible and distinct between acidic and neutral pH soils (Supplementary Figure 2a). Twenty of the 23 medium and high-quality MAGs recovered in ^{13}C-derived metagenomic libraries were from methanotrophic or non-methanotrophic methylotrophic populations and therefore consistent with targeting a methane-fuelled community using stable isotope probing (Supplementary Table 2). Nevertheless, comparison with equivalent ^{12}C incubations was performed as per standard practice with DNA-SIP experiments (Nicol and Prosser, 2011). Genomic DNA was recovered and purified from high buoyant density fractions (> 1.732 g ml^{-1}) from triplicate microcosms of both ^{12}C- and ^{13}C-CH_4 incubations. While recovered DNA from ^{12}C incubations was considered too low for metagenome sequencing, with DNA concentrations below the limit of detection in some

fractions, PCR amplification enabled characterisation of 16S rRNA gene-based community structures and comparison to the equivalent fractions from ^{13}C incubations.

Eleven families each contained >1% of annotated contigs from metagenome analysis of ^{13}C-enriched DNA from both soils, and they represented 47.4% and 85.3% of 16S rRNA gene amplicons in pH 4.5 and 7.5 samples, respectively, compared to 18.6 and 22.6% in the equivalent DNA fractions from ^{12}C-incubations. Annotated metagenome contig and 16S rRNA amplicon libraries generated from the same ^{13}C-enriched DNA both contained representatives of the same C1-utilising groups, although substantial differences were observed in relative abundance. For example, while the *Methylococcaceae* was the dominant family in pH 7.5 ^{13}C-enriched DNA using both approaches, it represented 23.4% (±0.5% s.e.) and 61.1% (±3.2% s.e.) of annotated contigs and 16S rRNA gene amplicons, respectively. In ^{13}C-enriched pH 4.5 DNA, the dominant family in the metagenomic libraries was the *Methylocystaceae*, representing 25.6% (±1.8% s.e.) of annotated contigs but only 0.02% (±0.006% s.e.) in the amplicon libraries, where the non-methanotrophic methylotrophic *Hyphomicrobiaceae* was the most abundant at 22.0% (±5.7% s.e.). Observed differences were therefore likely the result of several biases associated with different marker-gene and metagenomic sequencing approaches (McLaren et al., 2019). The overall community composition determined by 16S rRNA gene amplicon libraries were highly reproducible and clearly distinct between ^{12}C and ^{13}C incubations, confirming that communities analysed in metagenomic libraries were enriched in methane-derived ^{13}C (Supplementary Fig. 2c).

Predicted C1 metabolism in metagenome assembled genomes

Aerobic methylotrophic organisms utilise C1 compounds such as methane or methanol for both carbon and energy requirements (Chistoserdova et al., 2009). Methanotrophs are one group of methylotrophs that oxidise methane to formaldehyde via methanol, which is either assimilated for generating biomass or oxidised through to CO_2 to obtain energy and reductant. Non-methane oxidising methylotrophs, lacking methane monooxygenase, utilise methanol produced from other sources including that excreted from methanotrophs. Methane can therefore be directly and indirectly utilised by methylotrophic populations in natural communities.

To determine whether MAGs represented methanotrophic or non-methanotrophic methylotrophs, the presence of genes encoding methane monooxygenase (MMO), methanol dehydrogenases (MDH) and formate dehydrogenases (FDH) was determined after predicted

66 protein sequence annotation. Taxonomic assignment and methylotrophic characterisation were consistent with known traits of *Methylobacter*, *Methylocapsa*, *Methylosinus* and *Methylocystis*
68 strains, all of which possessed particulate methane monooxygenase (pMMO) and with the three *Methylosinus* MAGs also possessing soluble methane monooxygenase (sMMO). Eight MAGS
70 lacked genes encoding an MMO but possessed MDH and FDH confirming methylotrophic capability. This included representatives of previously recognised non-methanotrophic
72 methylotrophs including *Gemmatimonadales*, *Hyphomicrobium*, *Methylophilaceae*, *Methyloceanibacter* plus MAGs representative of the genera *Rudea* and *Herminiimonas*. One
74 MAG belonged to the class *Kiritimatiellae* of the Verrucomicrobiota. While this phylum contains known methylotrophs, no pathways for C1 metabolism were identified, potentially
76 due to the low (51.5%) estimated completeness.

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

80 Using assembled contigs >10 kbp, metagenomic viral contigs (mVCs) were predicted using two established tools. VirSorter (Roux et al., 2017) uses a database of viral genes plus analysis
82 of virus-like motifs, and DeepVirFinder (Ren et al., 2020) uses a *k*-mer based alignment-free approach, using viral genomes to train the prediction model. Both approaches provide different
84 levels of confidence with VirSorter categories 1, 2 and 3 representing 'most confident', 'likely' and 'possible' predictions, respectively, and DeepVirFinder also predicting likely (≥ 0.9, *p*-
86 value <0.05) and probable (≥ 0.7, *p*-value <0.05) viruses (Trubl et al., 2020).

We considered the matching of CRISPR array spacers with virus protospacers as the
88 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
90 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
92 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
94 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.

96

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

98 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

100 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 WIsH
102 (Galiez et al., 2017) 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
104 21 mVCs predicted to hosts via CRISPR spacer analysis were linked to a host using WIsH using a probability score ≤0.05 (Supplementary Table 8), all of which were consistent at the
106 family level. Of the 103 mVCs linked in vConTACT2.0 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
108 homologue-based prediction at the order and family levels, respectively. These analyses therefore indicated that while the *k*-mer based approach using WIsH was partially successful
110 in identifying correct linkages at the family level, it was not robust for identifying linkages with a high level of taxonomic resolution or confidence.

112

Gene-sharing networks of viral metagenomic contigs

114 Predicted hosts of mVCs from this study (using CRISPR and homologue-based approaches) were generally inconsistent with the known hosts of RefSeq viruses that were linked through
116 vConTACT2.0 gene-sharing network analysis (Jang et al., 2019). However, mVCs from this study placed into individual clusters all had the same predicted host (Supplementary Figure 4).
118 Initial networks used mVCs with a conservative host prediction only (i.e. a minimum of ≥ 5 homologues linked to one taxonomic family). With the exception of one cluster of 3 mVCs,
120 which were predicted to be linked to *Methylococcaceae* and non-methane oxidising methylotrophic *Hyphomicrobiaceae*, all individual networks were restricted to one family for
122 methanotrophs or one cluster of exclusively rhizobia-linked mVCs. Further analyses included all mVCs >10 kbp in this study (i.e. including those with < 5 host associated homologues) and
124 the same networks were identified i.e. mVCs linked within the same cluster all had homologues (i.e. with < 5 or ≥ 5) linking to them to same host family.

126 For *Methylocystaceae* mVCs, two separate but linked clusters were identified. The first contained CRISPR-linked *Methylocystaceae*-associated mVCs, which were recovered from
128 both pH 4.5 and 7.5 soil (although individual mVCs were restricted to one soil pH). A second cluster was dominated by mVCs from pH 4.5 soil only, indicating that most active
130 *Methylocystaceae* viruses belonged to one of two distinct lineages. There was a clear difference in the prediction of category-2 ('likely') and category-3 ('possible') mVCs associated with
132 these two clusters. While category-3 viruses are often excluded prior to analysis of soil viromes (Emerson *et al.*, 2019), CRISPR analysis demonstrated that one-third of linked mVCs were of

134 the lowest category of confidence. 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
136 exclusively of category-3 mVCs without further validation (e.g. without CRISPR spacer linkages) must be interpreted with caution. However, all major clusters identified through gene-
138 sharing network analysis contained a mixture of category-2 and -3 mVCs indicating that they represented groupings of genuine virus-derived genomes.

140 An intriguing finding was the linkage of host family-specific viruses from two different geographical regions (Scotland and Sweden) and contrasting soil types (agricultural loamy-
142 sand and permafrost peatland soils), from this study and that of Emerson et al. (2019), respectively. In the latter study, 13 of 1,907 mVCs were predicted to have a methanotroph host,
144 of which 9 were linked specifically to the *Methyocystaceae*. Intriguingly, 7 of these were also linked to a predicted *Methyocystaceae* mVC in our study, with one additional mVC predicted
146 to have a *Methylocapsa* host (i.e. belonging to the *Beijerinckiaceae* which is another methylotrophic family of the Rhizobiales) but also contained one predicted gene with a 'best
148 hit' match to a *Methylocystis* genome homologue. As soil pH is recognised as one of the dominant factors driving microbial community structures in soil (Bartram et al., 2014), it is
150 interesting to note that linked mVCs from both studies were also from acidic soils only.

152 **References**

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

156 Chistoserdova, L., Kalyuzhnaya, M.G. & Lidstrom, M.E. The Expanding World of Methylotrophic Metabolism. *Annu Rev Microbiol.* 63, 477–499 (2009).

158

Emerson, J.B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat.*
160 *Microbiol*. 3, 870-880 (2018).

162 Galiez, C, Siebert, M., Enault, F., Vincent, J. & Söding, J. WisH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33, 3113-3114 (2017).

164 Jang, H.B. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol*. **37**, 632-639 (2019).

166 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).

168

McLaren M.R., Willis, A.D. & Callahan, B.J. Consistent and correctable bias in metagenomic
170 sequencing experiments. *eLife* 8, e46923 (2019).

172 Ren, J. et al. Identifying viruses from metagenomic data by deep learning. *Quant. Biol.* 8, 64-77 (2020).

174

Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
176 microbial genomic data. *PeerJ* 3, e985 (2015).

178 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 Viromics. *Soil*
180 *Syst.* 4, 23 (2020).