Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing

Rebecca A. Daly[1], Simon Roux[2], Mikayla A. Borton[3], David M. Morgan[4], Michael D. Johnston[4], Anne E. Booker[1], David W. Hoyt[5], Tea Meulia[6], Richard A. Wolfe[1], Andrea J. Hanson[7,8], Paula J. Mouser[7,8], Joseph D. Moore[9], Kenneth Wunch[10], Matthew B. Sullivan[1,7], Kelly C. Wrighton[1], and Michael J. Wilkins[1,4*]

[1]Department of Microbiology, Ohio State University, Columbus, OH, 43210
[2]Joint Genome Institute, Walnut Creek, CA, 94598
[3]Environmental Sciences Graduate Program, Ohio State University, Columbus, OH, 43210
[4]School of Earth Sciences, Ohio State University, Columbus, OH, 43210
[5]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, 99350
[6] Molecular and Cellular Imaging Center, Ohio State University, Wooster, OH, 44691
[7]Department of Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH, 43210
[8]Department of Civil and Environmental Engineering, University of New Hampshire, Durham, NH, 03824
[9]Dow Microbial Control, Collegeville, PA 19087
[10]Dow Microbial Control, Houston, TX 77077

[*]Corresponding author:

Michael J. Wilkins
The Ohio State University
School of Earth Sciences
Department of Microbiology
Mendenhall Laboratory
125 S. Oval Mall
Columbus, OH 43210
mjwilkins@gmail.com

Summary

The deep terrestrial biosphere harbors a significant fraction of earth's biomass and remains understudied compared to other ecosystems. Deep biosphere life primarily consists of bacteria and archaea, yet knowledge of their co-occurring viruses is poor. Here we temporally catalogued viral diversity from five deep terrestrial subsurface locations (hydraulically fractured wells), examined virus-host interaction dynamics, and experimentally assessed metabolites from cell lysis to better understand viral roles in this ecosystem. We uncovered high viral diversity, rivaling that of peatland soil ecosystems, despite low host diversity. Many viral OTUs were predicted to infect *Halanaerobium*, the dominant microbe in these ecosystems. Examination of CRISPR-Cas spacers elucidated lineage-specific virus-host dynamics suggesting active *in situ* viral predation of *Halanaerobium*. These dynamics indicate repeated viral encounters and changing viral host range across temporally and geographically distinct shale formations. Laboratory experiments showed that prophage-induced *Halanaerobium* lysis releases intracellular metabolites that can sustain key fermentative metabolisms, supporting the persistence of microorganisms in this ecosystem. Together these findings suggest that diverse and active viral populations play critical roles in driving strain-level microbial community development and resource turnover within this deep terrestrial subsurface ecosystem.

Main Text

Bacteria and their viruses are two of the most abundant and genetically diverse entities on Earth[1,2], and are inferred to play central roles in the biochemical functioning of every ecosystem studied to date. Viruses affect community turnover and resource availability via a range of interactions with their hosts. Through predation, viruses exert both top-down controls by decreasing bacterial densities and bottom-up controls by lysing cells and releasing labile cellular contents[3,4]. During infection, viruses also influence key cellular processes, such as metabolism, division, CRISPR-Cas immunity, motility and regulation[5,6]. Thus, while viral studies in nature remain woefully behind those of their microbial counterparts it is clear that viruses impact ecosystem processes in diverse ways.

Hydraulically fractured shales represent an opportunity to investigate viral effects on ecosystem functioning in the deep terrestrial subsurface. Hydraulic fracturing involves the high-pressure injection of water and chemical additives into a shale formation, generating fracture networks that release oil and gas. During this process, microorganisms from the surface are concomitantly injected with water and chemicals and colonize the fractured shale. This ecosystem is characterized by high temperatures, pressures and salinities, constraining the diversity of microorganisms that inhabit it. Low genus-level diversity contributes to genomic tractability, and is coupled with relatively fast microbial growth rates that offer an opportunity to study *in situ* processes that might occur over hundreds or thousands of years in other deep subsurface ecosystems.

To date, viral impacts on ecosystem functioning have primarily been explored in marine, soil, host-based systems, and in the shallow terrestrial subsurface[7-12]. Thus, although the deep biosphere represents the largest biotope on earth[2], our knowledge of viruses and

their ecological role in the deep terrestrial subsurface is largely unknown[13]. Previously, we examined the microbial and viral diversity in a single well in the Marcellus shale formation and found evidences of ongoing viral-host interactions for the five bacterial and archaeal genera persisting over 300 days after hydraulic fracturing. These findings provided the first evidence for active viral predation in the deep terrestrial biosphere[14]. Here we expand the dataset to five hydraulically fractured wells (including a well from our previous study) to describe viral-*Halanaerobium* interactions across hydraulically fractured wells and the most extensive census of viral populations in the deep (>1000 m) terrestrial subsurface to date.

Results and Discussion

Construction of a genomic Halanaerobium *database from fractured shales*

Microbial communities in shale wells are characterized by progressively lower diversity after the hydraulic fracturing process, eventually converging to a core terminal community dominated by anaerobic *Halanaerobium* spp.[14-18]. These microorganisms are of broad interest given their ability to degrade input fluid chemistries[16], form biofilms and contribute to biofouling[15,17], and generate toxic sulfide[15,16]. In order to understand *Halanaerobium*-virus interactions in this ecosystem, we created a database of 21 *Halanaerobium* metagenome-assembled genomes (MAGs) and isolate genomes (Supplementary Table 1). Five MAGs were from the Marcellus-1 and Utica-2 wells, including two from the Utica-2 well glycine betaine enrichment. In addition, eleven *Halanaerobium* spp. were isolated from the Utica-2 and Utica-3 wells, and we also included publicly available described *Halanaerobium* isolate genomes including two genomes from hypersaline oil reserve brines[19,20] and three genomes isolated from surface environments[21-23].

Diversity of viruses in fractured shales rivals that of terrestrial and marine ecosystems

Viruses can be examined in microbial size-fraction metagenomes due to captured lytic viral infections, integrated proviruses, and free lytic particles co-sampled on produced fluid filters prior to DNA extraction. Viral sequences were identified using a combination of the presence of viral "hallmark" genes, Pfam gene depletion, a high-degree of uncharacterized proteins, a low degree of strand switching, and an enrichment in short genes compared to bacterial and archaeal genes[24]. We identified 1,838 predicted viral populations (viral operational taxonomic units, viral OTUs) ≥5 kb length, which are suggested to represent approximately species-level taxonomy by studying gene flow and selection in cyanophage populations[25]. Host diversity in hydraulically fractured shale produced fluids (Shannon's diversity *H'*=0.61 ±0.49 s.d.) is lower than most surface biomes where *H'* can exceed 10.0 (*e.g.,* grassland soils)[26] and is typically confined to less than ten microbial OTUs ~50 days after the initial hydraulic fracturing[18]. Surprisingly, viral shale diversity was disproportionately higher, with a viral OTU accumulation curve from the five shale wells rivaling a similar curve generated from soils (Fig. 1a)[9], an ecosystem typically considered to be the model for high microbial diversity.

Because viruses lack a universal barcode gene for inferring taxonomy, we instead employed a genome-based network analysis of shared protein content[27]. Specifically, the

1,838 fracking viral OTUs were analyzed alongside publicly available, curated bacterial and archaeal viral genome sequences[28] and metagenomes[29]. The resulting network of viral OTUs and reference sequences were clustered into viral clusters (VCs), which taxonomically approximate a viral genus[27] (Fig. 1b, Supplementary Data 2). Viral OTUs clustering with ICTV-classified reference sequences were associated nearly exclusively (94.9%) with genera in the dsDNA viral order Caudovirales (Fig 1b), representing only 34.8% of the viral OTUs. Within the Caudovirales 45.8% were in the family Myoviridae, 30.3% in the family Siphoviridae and 23.9% in the family Podoviridae. The majority of shale viral OTUs could not be assigned to ICTV taxonomic groups (46%), yet these could nevertheless be classified into 156 previously undescribed candidate viral genera, i.e., genus-level groups which contained exclusively viral OTUs from fractured shales. This finding highlights the uniqueness of viral genomes unearthed in this study, comparable to the previously undescribed candidate genera found in the Global Oceans Viromes dataset (658 previously undescribed candidate genera from 104 viral-targeted samples)[28], and in soils (451 previously undescribed candidate genera from 214 microbial-targeted samples)[9], ecosystems with significantly higher host diversity than fractured shales.

Despite this strong shale viral signature, no single viral OTU was shared across all four sites, indicating that the high viral diversity observed in hydraulically fractured shales may be site-specific (Fig. 1c). This suggests a potentially strong founder effect or very rapid diversification processes in these wells, which are essentially closed ecosystems after the hydraulic fracturing process. While the sequences identified in this study are likely not indigenous, further study is needed to definitively determine their original source. Viral OTUs were then further classified into viral genera, using a network analytic based on shared protein content with viruses having official viral taxonomy by the International Committee on Taxonomy of Viruses (ICTV)[28,30]. Despite the lack of shared viral OTUs across sites, 17 genera (6.5%) were shared across sites, including two genera within the *Siphoviridae*, one genus within the *Myoviridae*, and 14 previously undescribed shale genera (Supplementary Fig. 1). This observed pattern is likely due to the different levels of taxonomic resolutions examined; species-level taxonomy captures specific virus-host interaction dynamics, whereas genus-level taxonomy typically represents how the virus replicates, packages, and transports its genome.

Linking viruses to* Halanaerobium *hosts

Given the prevalence and dominance of *Halanaerobium* reported by many studies across geographically and geologically distinct shales[14,16-18,31-34], we next identified the viruses associated with *Halanaerobium* hosts and examined their temporal abundance patterns in detail. Two methods were used to identify viruses as having *Halanaerobium* host(s). First, using an oligonucleotide frequency dissimilarity measure[35] we determined that 21.6% of viral OTUs were putatively associated with at least one of the *Halanaerobium* genomes included in this study (Supplementary Data 2), consistent with the dominance of this genus in the community. Second, using a more stringent method, we matched CRISPR-Cas spacer sequences from *Halanaerobium* MAGs and isolate genomes to viral protospacers in 32 viral OTUs (Supplementary Data 2).

Next, we contrasted the relative abundances of *Halanaerobium*-associated viral OTUs with temporal patterns of *Halanaerobium* 16S rRNA gene relative abundances across the five wells (Fig. 2). This revealed nearly identical temporal patterns between viral and *Halanaerobium* relative abundances, similar to the correlation observed with cyanophage and *Synechococcus* spp.[36]. Because this pattern could be driven by integrated prophage in *Halanaerobium* genomes, we examined the fraction of predicted viral sequences with a putative lysogenic lifestyle in input and produced fluids. 50 of the predicted viral sequences were identified as integrated prophage, and an additional 55 viral sequences were identified as having a putative lysogenic lifestyle based on gene annotations such as integrase and excisionase genes. We found that predicted lysogenic viruses made up a significantly higher percentage of the viral pool in produced fluids compared to input fluids (produced fluids mean=24.9%, s.d. ± 10.4%; input fluids mean=4.41%, s.d. ± 4.0%). (Supplementary Fig. 2). In the temporally-resolved datasets, especially notable were two separate crashes in *Halanaerobium* relative abundances that occurred in both Utica wells (U-2, U-3) between days 75-105 and were coordinated to changes in viral abundances. This tight coupling between viral and host relative abundances (Spearman's rho = 0.828, $p<0.001$) suggests that *Halanaerobium* in shale fluids are surviving despite predation from free viruses and induction of prophage. Given the central role that *Halanaerobium* plays in the functioning of this ecosystem[14-17], viral infection of *Halanaerobium* populations would likely impact key ecosystem processes including carbon cycling, biofouling, and sulfidogenesis.

We next leveraged our most highly resolved time series of metagenomic data supplemented with cultivation and genomic sequencing from the Utica-2 well. Instead of examining the aggregate abundances of *Halanaerobium* spp. and viral OTUs, we examined a genome-resolved view of *Halanaerobium* and viral dynamics (Fig 3, Supplementary Data 2, Supplementary Data 3). *Halanaerobium* spp. have high average nucleotide identity (ANI) values across shared genomic regions, ranging from 99.76-100.00% (Supplementary Table 2). For example, comparisons of the *Halanaerobium* isolates WG6 and WG7 reveals that they share identical genomes, including the same prophage, with the exception that *Halanaerobium* WG6 has a type I CRISPR-Cas system and 15 spacers that the WG7 isolate lacks. This taxonomic resolution allows us to uncover viral-host dynamics at the strain level.

Halanaerobium isolates and MAGs show large temporal changes in relative abundance between days 86 and 105 in the Utica-2 dataset (Figure 3a). The MAG genome 5U2 shows a unique pattern, dominating during this early period and declining below detection during the remainder of the sampling period. Other isolate and MAG genomes show a similar response between days 106-175, although less dramatic than in earlier time points, and show relatively stable abundances between days 176-302. Due to the stable geochemical conditions in these wells[18], lack of external inputs after the hydraulic fracturing process, and absence of other bacterial predators (*e.g.,* protozoa, *Bdellovibrio* spp.), we infer that these changes in *Halanaerobium* strain abundances could be due to viral predation. A decrease in the abundance of one *Halanaerobium* genotype creates niche occupancy for another *Halanaerobium* strain to fill. These abundance patterns are similar to those of *Halanaerobium* associated viral OTUs (Figure 3b), with those detected

early in the sampling time period (86-105 days) decreasing to negligible abundances in latter time points, when other viral OTUs become abundant. Additionally, these viral OTUs show the most dynamic abundance behavior during days 106-175, with relatively stable abundances during days 176-302, similar to the *Halanaerobium* genotypes.

The few studies that have resolved the dynamics of viruses and their hosts *in vivo* at fine temporal scales (*i.e.* days to weeks) have either focused on the human microbiome[37] or relied on *in vitro* assays using a limited number of hosts[38]. The complex viral-host abundance patterns observed in this ecosystem are a likely result of both the viral requirement of host cells for growth, and the fact that viruses are the only predator in hydraulically fractured shales. This predation pressure enhances natural selection in ecosystems, as acquisition of immunity by a *Halanaerobium* strain could lead to near extinction of the virus. Conversely, increased virulence risks extinction of the host population. This evolutionary relationship is driven by co-adaptation that supports the reproduction and very existence of both virus and host[39]. Here, within a single well, highlighting the dynamics of host genomes within a single genus and their associated viruses, we see an intricate pattern of viral-host interactions. These dynamics suggest that complex mechanisms affect host defense and viral counter-defense in this ecosystem[40].

Viruses target Halanaerobium strains during the hydraulic fracturing process

Many bacteria rely on innate and adaptive immune systems to defend against viral invasion. The CRISPR-Cas adaptive immune system provides sequence-specific protection against invading DNA through incorporation of fragments of foreign DNA from coexisting infective viruses (protospacers) into a CRISPR array as recently acquired spacers. The spacers provide an evolving, genetic memory for targeted defense against subsequent invasions by the virus. Viruses can evade or subvert CRISPR-Cas systems by accumulating mutations in the protospacers or its adjacent motif; bacteria can counter viral evasion by accumulating spacers from the viral genome.

To further elucidate *Halanaerobium*–viral interactions in this ecosystem, we analyzed *Halanaerobium* isolate genomes and MAGs for the presence of CRISPR-Cas arrays. CRISPR-Cas systems were detected in 14 of the 21 *Halanaerobium* genomes and were dominated by Type-I CRISPR-Cas (Fig. 4, Supplementary Table 1). Next, we sought to leverage the archival nature of CRISPR-Cas array spacers in the microbial genomes to reveal the details of specific virus-host interactions. To date few studies have successfully linked CRISPR-Cas array spacers in host genome to protospacers in viral genomes in nature. One example is a study that analyzed CRISPR-Cas arrays from acid mine drainage over a 5-year period in order to elucidate the genetic interplay between *Leptospirillum* and the AMDV1 virus[41]. In a hypersaline lake ecosystem study, 20 viral populations were matched to spacers in metagenomic samples from a database of 140 viral populations detected across 17 samples[42]. Here, we detected a total of 649 spacers within the *Halanaerobium* genomes and were able to match 151 spacers (141 perfect matches, 10 with a single bp mismatch) to protospacers in 32 viral OTUs. No links were found to viruses in the RefSeq database and links were only detected between *Halanaerobium* and viral OTUs sampled from oil and gas wells. These viral and host populations may be endemic to the deep terrestrial subsurface in oil and gas reservoirs,

although further metagenomic sampling of wells for energy production, including the study of their viral communities, are necessary to determine if this is the case.

We queried CRISPR-Cas array composition from *Halanaerobium* MAGs across temporal metagenomic samples to determine if spacers were being acquired during the sampling period. We found that *Halanaerobium* 5-U2 and *Halanaerobium* 6-U2 had acquired one and eight spacers respectively over 216 days (Supplementary Fig. 3). In an earlier study, we found one *Halanaerobium* MAG, *Halanaerobium* 1-M1, had acquired two spacers over a 247 day sampling of this genome[14]. These spacers were not incorporated at the leader end of the CRISPR-Cas array as is usually observed. While this may be an artifact of an assembly error in the array, ectopic spacer incorporation has previously been observed in *Streptococcus thermophilus*[43], and may suggest a mutation in the array leader sequence. In all three cases, none of the recently acquired spacers matched sequences in our viral database. Recentlyy acquired spacers may represent low abundance viruses whose sequences were not assembled, and may have remained at low abundance due to the immunity conferred to *Halanaerobium* by recent spacer incorporation. However, the combination of observed virus dynamics and the acquisition of spacers by *Halanaerobium* genotypes during the sampling period provide additional lines of evidence that viral predation and host immunity is active in these ecosystems.

Analysis of the viral-host network (Fig. 4) established that 66% of the nine genomes with detected links had multiple spacer links to the same viral OTUs[44,45]. Strikingly, a CRISPR-Cas array in the *Halanaerobium* 6-U2 MAG with a total 43 spacers had 20 links to the same viral sequence. This may be because mutations in a viral protospacer sequence would require a *Halanaerobium* host to acquire new spacer sequences in order to maintain immunity to the mutated virus, or could reflect how different members of a population might incorporate different spacers when challenged with the same virus. Additionally, more than half (57.1%) of the viral OTUs with matching protospacers were linked to multiple *Halanaerobium* genomes. These data suggest that many of these viral OTUs have a broad host range, infecting multiple *Halanaerobium* hosts. In addition, these data imply long-term coexistence of *Halanaerobium* with the viruses detected in this study

We specifically then asked whether any *Halanaerobium* genome had acquired CRISPR-Cas mediated immunity to a prophage carried by another *Halanaerobium* strain. In our viral genomic dataset, we identified 50 viruses with a putative lysogenic lifestyle and three integrated prophages from *Halanaerobium* isolate genomes (*H. salsuginis*, *H. congolense* WG7, *H. congolense* WG8). Eight of the 35 viral populations with CRISPR-Cas links were identified as lysogenic either by flanking bacterial genome sequences or by the presence of genes such as integrases. Viral cluster 68 was of particular interest as the two reference sequences used to assign taxonomy to the VC were from induction experiments and four of the shale-derived viruses from VC 68 were identified as integrated prophage (Supplementary Table 3). Members of this viral genus showed conserved gene order and shared homologous sequences (Supplementary Fig. 4). Viruses in VC68 were linked to three MAGS and two isolate *Halanaerobium* genomes by CRISPR spacer matches. The *Halanaerobium* 1M1 MAG had three CRISPR-Cas links to

the prophage that was identified in the genome of *H. salsuginis*, isolated from a traditional oil well in Oklahoma[20]. Additionally, the *H. salsuginis* genome has a CRISPR-Cas array with two spacers matching viruses in this dataset of Utica and Marcellus formation-derived viruses. This implies that while viral OTUs appear to be site-specific (Figure 1b), geographically distinct *Halanaerobium* in oil and gas wells have seen similar viruses (genus level). Together these data suggest that lysogenic viruses may be an important pool of viral diversity in hydraulically fractured wells.

Environmental stressors induce Halanaerobium prophage and cause release of labile substrates into extracellular environment

We propose that viral predation may not only exert top down control on shale microbial communities, but may also provide an important resource to the persisting microbial community in this closed ecosystem, via release of labile intracellular contents into the environment. To test this hypothesis, we evaluated our *Halanaerobium* isolates for evidence of prophage induction upon the addition of chemical stressors. We conducted a phage-induction experiment with the isolate *H. congolense* WG8 by treating cultures with mitomycin C, and two environmentally relevant stressors: the organic acid succinate and the heavy metal copper chloride. Eighteen hours after copper chloride and succinate treatments, cultures showed a decrease in cell density of 58% and 54% respectively, indicating cell lysis. After induction with succinate, viral particles were enriched 40x relative to pre-induction counts and the burst size was calculated to be ~61 viruses per lytic event. Transmission electron microscopy analysis of post-induction succinate cultures showed tail-less viral particles of 27.5 ±3.8nm in diameter and *Halanaerobium* cell lysis concurrent with viral particle release (Figure 5b, c). Interestingly, while most work in viral ecology has focused on tailed dsDNA viruses so far (*Caudovirales*), the importance of non-tailed viruses is beginning to be recognized[46,47].

Upon viral lysis, we purified and sequenced the post-induction viral DNA to reveal a single region of the WG8 genome that was enriched in the extracellular fraction. This ~45.8 kb region is characterized by genes annotated as integrase, excisionase and hypothetical proteins; is flanked by regions rich in transposases, but lacks hallmark viral genes (including putative capsid genes when searching against recently described non-tailed viruses[47]). Collectively, these data likely explain why it was not identified using traditional *de novo* bioinformatic predictions (Figure 5d). One gene was annotated as a superinfection exclusion protein; a type of protein that is known to prevent secondary viral infections to their hosts while integrated[48]. Given that we found no genomic evidence of a CRISPR-Cas system in the *H. congolense* WG8 genome, it raises the question if conferring superinfection exclusion to the host is a trade-off for harboring the prophage and the subsequent potential for cell lysis.

Collectively these data show that viral lysis of microbial biomass in this ecosystem likely exerts top-down controls on host populations. Further, lysis also exerts bottom-up controls by releasing compounds from the lysed cells that can be consumed by other populations. Extracellular metabolites were measured in culture media before and after the induction of the prophage in *H. congolense* WG8, revealing increases in the concentrations of seven compounds, primarily amino acids and organic acids after viral

lysis (Figure 5e). While members of the fractured shale community encode the genomic capacity to utilize these compounds as carbon, sulfur, and nitrogen sources, the use of alanine and valine as electron donors in Stickland reactions is thought to be especially important for sustaining *Halanaerobium* long after the degradation of more labile input material from the initial hydraulic fracturing event[14,49]. Our results here specifically indicate that microbial cell lysis from viral infections has the potential to support a network of interdependent microorganisms, allowing them to persist in this ecosystem. More generally, the impact of viruses on nutrient fluxes is progressively recognized, and it has been suggested that viral predation needs to be taken into account when modeling microbial food webs and ecosystem processes[3,50]. Our results contribute to the growing body of research suggesting that viral ecological data are important for the development of predictive understanding of microbial ecosystem functioning.

Conclusion

Here we describe a genome-resolved (via metagenomic bins and sequenced isolates) survey of the dominant bacterium, *Halanaerobium* spp., across hydraulically fractured wells and the most extensive census of viral OTUs in the deep (>1000 m) terrestrial subsurface to date. The viral diversity in this ecosystem is surprisingly high given the low host diversity, and rivals the diversity observed in terrestrial and marine ecosystems. Our results provide evidence for extensive viral predation and host adaptive immunity at the individual genome level as top-down controls on *Halanaerobium* populations. Viral-host interaction patterns suggest the long-term coexistence of *Halanaerobium* hosts and viruses. Through integrated analyses of metagenomic data and laboratory experiments, we demonstrate that active viruses also exert bottom-up controls that ultimately support microbial persistence in the fractured shales. These findings have important ramifications for understanding the dominance and persistence of multiple closely related *Halanaerobium* in this economically important ecosystem. Our results highlight the need to examine the nature of viral-host interactions across different ecosystems at the genome-resolved level.

Much of the microbial genetic diversity we observe in ecosystems is likely an aggregate of microdiverse strains grouped by similar 16S rRNA gene sequences. Making inferences based on aggregate responses obscures the complexity of viral-host interactions at relevant strain or population levels, and blurs the effects of viral predation on microbial resilience, resistance, and ecosystem function. Furthermore, much of our knowledge derives from laboratory model systems that underrepresent the diversity intrinsic in natural ecosystems. Hydraulically fractured shales dominated by *Halanaerobium* and containing abundant, diverse viruses may be a tractable, model system for studying viral-host interactions and the effects on both host microdiversity and resulting community assembly.

Methods

Sample collection and DNA extraction

Hydraulic fracturing input fluids and shale produced fluids were collected from wellheads and oil-gas-water separators in sterile bottles with no headspace (1-2 L). When

fluids were collected from the separator tanks, the separator tanks were flushed immediately prior to sample collection to minimize community changes due to incubation in the separator. Flow rates ranged from ~400,000L per day at early time points to ~170,000 L per day at later time points, with a separator capacity of ~5,500 L.These fluids were collected from five wells in the Marcellus and Utica-Pt. Pleasant shale formations in Ohio (n=2, Utica formation), West Virginia (n=2, Marcellus formation) and Pennsylvania (n=1, Marcellus formation). The two wells in West Virginia were located on the same wellpad yet subjected to different hydraulic fracturing methods. Also included in analyses for viral contig identification was a sample from a laboratory-based glycine betaine enrichment experiment using fluids from the Utica-2 well (day 96 post-hydraulic fracturing)[14]. Fluids were filtered on 0.22 μm pore size polyethersulfone (PES) filters (Millipore, Fisher Scientific). Produced fluids contained high concentrations of ferrous iron (100-180 mg/L). During the filtering process the oxidizing iron resulted in a natural analog of the chemical flocculation protocol used to concentrate free viral particles[51] (Supplementary Fig. 5). Thus, the viruses sampled in this study are likely a combination of free viral particles, integrated proviruses and active, replicating viruses. Total nucleic acids were extracted from filters for the Marcellus well located in Pennsylvania as previously described[14]. For the other four wells, nucleic acids were extracted using Lysis Buffer I[52], purified with two phenol-chloroform and one chloroform-isoamyl alcohol extraction, and precipitated with NaCl and ice-cold ethanol. Nucleic acids were stored at -20°C until sequencing.

In this study we sampled fluids from five hydraulically fractured wells in the Utica and Marcellus shale formations in Pennsylvania, Ohio and West Virginia, U.S.A. These samples included seven input fluids that were injected into the wells during the hydraulic fracturing process and 33 produced fluid samples collected from the wells during oil and gas production (Supplementary Table 4, Supplementary Data 1). A glycine betaine enrichment using fluid from the Utica-2 well (96 days after hydraulic fracturing) was also included in these analyses in order to maximize the number of *Halanaerobium* genomes in the database[14]. The best longitudinal sampling was obtained from the Utica-2 well with sampling initially occurring daily when the well first started oil and gas production and gradually decreasing in frequency through 302 days after hydraulic fracturing. Collectively these wells cover the major hydraulic fractured shale formations across the Appalachian Basin.

Metagenomic sequencing, assembly and annotation

Libraries were prepared using the Nextera XT Library System according to manufacturer's instructions. Sequencing adapters were ligated and library fragments were amplified with 12-15 cycles of PCR before quantification with the KAPA Biosystems next-generation sequencing library quantitative PCR kit. Following library preparation with a TruSeq paired-end cluster kit (v4), sequencing was performed on the Illumina HiSeq 2500 platform with HiSeq TruSeq SBS sequencing kits (v4) following a 2x150 bp indexed run recipe. Fastq files were generated with CASSAVA 1.8.2. Fastq files were trimmed from the 5' and 3' ends with Sickle (https://github.com/najoshi/sickle) and each sample was assembled individually with IDBA-UD using default parameters.

Metagenome statistics including the amount of sequencing are detailed in Supplementary Table 1. Scaffolds ≥ 5.0 kb were included in subsequent analyses. Scaffolds were annotated as previously described[14].

Halanaerobium *genomic analyses*

Halanaerobium metagenome-assembled genomes (MAGs) were obtained using manual binning by a combination of phylogenetic signal, coverage and GC content. For each bin, genome completion was estimated based on the presence of core gene sets for Bacteria (31 genes) and Archaea (104 genes) using Amphora2[53]. Genome completion is reported in Supplementary Table 2. *Halanaerobium* bacteria were isolated as previously described[15]. Near-full-length ribosomal 16S rRNA gene sequences were reconstructed from unassembled Illumina reads using EMIRGE[54]. To reconstruct 16S rRNA gene sequences we followed the protocol with trimmed paired-end reads where both reads with at least 20 nucleotides used as inputs and 50 iterations. EMIRGE sequences were checked for chimera before phylogenetic gene analyses. The relative abundance of *Halanaerobium* genomes was calculated using read-mapping with Bowtie2[55]. Reads from each sample were mapped to genomes using multimapping zero mismatches. To be included in the abundance analyses we required that at least 90% of each genome's contigs were covered 5x by reads over a minimum of 75% of the contig length. Genome abundances were obtained by dividing the number of reads mapped by the length of the genome (normalizing for differences in sequence length) and summing the values in each sample. Then the length normalized values were divided by the summed value in each sample (normalizing for the amount of reads mapped in each sample) (Fig. 3a). A concatenated ribosomal tree (Fig. 4) was constructed for *Halanaerobium* MAGs and *Halanaerobium* isolate genomes using 16 ribosomal proteins chosen as single-copy phylogenetic marker genes (RpL2, RpL3, RpL4, RpL5, RpL6, RpL14, RpL15, RpL16, RpL18, RpL22, RpL24, RpS3, RpS8, RpS10, RpS17, and RpS19). Each individual protein data set was aligned with MUSCLE 3.9.31 and manually curated to remove end gaps. Alignments were concatenated to form a 16-gene alignment and run through ProtPipeliner. Alignments were curated with minimal editing by Gblocks, with the LG model selected by ProtTest 3.4, and a maximum-likelihood phylogeny for the concatenated alignment was generated with RAxML version 8.3.1 using 100 bootstrap replicates.

Viral analyses

Viral sequences were identified in all assembled metagenomes using VirSorter[24] hosted on the CyVerse discovery environment. VirSorter was run with default parameters using the 'virome' database and viral sequences with category 1 and 2 status were retained. Prophage in *Halanerobium* genomes were identified using the same methods. Viral sequences were clustered into viral OTUs using the 'ClusterGenomes' (v 1.1.3) app in CyVerse using the parameters 95% average nucleotide identity and 80% alignment fraction of the smallest contig[28]. In order to compare viral OTU accumulation curves between soils[56] and the shale wells in this study, only viral sequences ≥ 10 kb or circular were included in the analysis (Fig. 1a). Additionally, because the soil study referenced here included 236 metagenomes and seven viromes the species accumulation curve was truncated at 41 samples (the number of samples in the shale dataset). In order to detect

low-abundance viruses that did not assemble in each sample, and to calculate viral relative abundance (abundance within the viral community), we used a read-mapping with Bowtie2[55]. Reads from each sample were mapped to viral OTU sequences using multimapping and zero mismatches. To be included in the viral abundance analyses we required a depth of 5x coverage across at least 75% of the viral sequence. Viral OTU abundances were obtained by dividing reads by the length of the viral sequence (normalizing for differences in sequence length) and summing the values in each sample. Then the length normalized values were divided by the summed value in each sample (normalizing for the amount of reads mapped in each sample). A network-based protein classification was used to taxonomically place the shale-derived viral sequences in the context of known viruses according to the ICTV[28,30]. Predicted proteins from viral contigs were clustered with predicted proteins from viruses in the NCBI RefSeq database (v75, July 2016) based on all-vs-all BLASTp search with an *E*-value of $1e^{-3}$ and protein clusters (PCs) were defined with Markov clustering algorithm (MCL) as previously described[28,30] and processed using vContact[27]. The stringency of the similarity score was evaluated through 1,000 randomizations by permuting PCs or singletons (proteins without significant shared similarity to other protein sequences) within pairs of sequences having a significance score ≤1 (negative control)[57]. Subsequently pairs of sequences with a similarity score >1 were clustered into viral clusters (VCs, ~genus level[24] with MCL using an inflation value of 2, as previously described[28]. The resulting network was visualized with Cytoscape software (version 3.5.1)[58] using an edge-weighted spring embedded model. Reference sequences that co-clustered with the shale-derived viral sequences were used to predict viral taxonomy with a last common ancestor approach. If the taxonomy of reference genomes within a VC differed, majority rule was used. If VCs exclusively contained shale-derived viral sequences from this study, the VC was termed "previously undescribed." Genome synteny plots of viral OTUs in VC68 (Supplementary Fig. 3) were performed using EasyFig (v2.1)[59] based on a full-genome BLASTn search.

Halanaerobium *CRISPR-Cas analyses and viral host prediction*

The CRISPR Recognition Tool plugin (CRT, version 1.2) in Geneious was used to identify CRISPR arrays in *Halanaerobium* MAGs and isolates. To identify matches between viral protospacers and *Halanaerobium* CRISPR-Cas array spacers we used BLASTn with an *E*-value cutoff of $1e^{-5}$. All matches were manually confirmed by aligning sequences in Geneious; one bp mismatch was allowed. Links between viral sequences and *Halanaerobium* were used to construct a network (Fig. 4). *Halanaerobium* CRISPR-Cas systems were classified by manually examining the CRISPR-Cas proteins of annotated contigs[60]. If a *Halanaerobium* genome had a spacer link to a viral protospacer, the host of that viral sequence was subsequently classified as *Halanaerobium*. In addition to host prediction based on CRISPR analyses, we also used the oligonucleotide frequency dissimilarity (VirHostMatcher) measure[35]. The host genomes used as input to VirHostMatcher included all *Halanaerobium* genomes in this study, other Firmicutes such as *Dethiosulftibacter* spp., *Frackibacter* spp., *Orenia* spp., and other core members of hydraulically fractured well communities[18] including the Archaea *Methanohalophilus* spp., *Methanolobus* and *Thermococcus* spp. We determined which dissimilarity measure option to use based on agreement to our CRISPR-Cas links;

25 of the 32 viruses with *Halanaerobium* CRISPR-Cas links agreed with the d_2^* measure and those data were used to further predict viral-*Halanaerobium* associations.

Halanaerobium congolense WG8 *prophage induction*

Halanaerobium congolense WG8 were grown in anaerobic, defined saltwater liquid medium as described in Booker *et. al.* 2017 at 40°C. After 72 hours of growth, mitomycin C (final concentration 0.5 μg/mL), succinate (100 mM), copper (II) chloride (10 mM) or anaerobic water (control) were added to triplicate incubations for each treatment. Samples of media were collected for metabolites immediately prior to addition of induction agents and 18 hours post-induction, filtered and sent to the Pacific Northwest National Laboratory for NMR analysis. Samples were diluted by 10% (vol/vol) with 5mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 as an internal standard. All NMR spectra were collected using a Varian Direct Drive 600 MHz NMR spectrometer equipped with a 5 mm triple resonance salt-tolerant cold probe. The 1D ^{1}H NMR spectra of all samples were processed, assigned and analyzed using Chenomx NMR suite 8.3 with quantification based on spectral intensities relative to the internal standard. Metabolites present in each sample were determined by matching the chemical shift, J-coupling and intensity information of experimental NMR signals against the NMR signals of standard metabolites in the Chenomx library. The 1D ^{1}H spectra were collected following standard Chenomx data collection guidelines[61], using a 1D nuclear Overhauser effect spectroscopy (NOESY) with presaturation (TNNOESY) experiment with a 100-msec mixing time, a 1.5s presaturation pulse, 65,536 complex points, a 12 ppm spectral width, an 4s acquisition time and at least 512 scans at 298K. Metabolite data is reported in Supplementary Data 2. Viral particles were enumerated by epifluorescence microscopy according Thurber *et al*. 2009[62]. Cells were stained with 2.5x SYBR gold solution for 10 minutes in the dark and counted on a Nikon ECLIPSE CiL LED microscope. Induced burst size (BZ_I) was calculated using the following equation: $BZ_I = (V_I - V_C/B_C - B_I)$ where V_I is the number of viruses in the induced sample and V_C is the number of viruses counted in the controls. B_C and B_I are the number of bacterial cells counted in the controls and induced samples, respectively[63]. Viral particles were purified using CsCl gradient density centrifugation and collected from the 1.3 g/mL layer. Viral DNA was extracted according to Thurber *et al*. 2009[62] and purified viral DNA was sequenced at the Joint Genome Institute. The post-induction viral sequencing was assembled using methods described above. The putative prophage was identified based on differential sequence coverage (putative prophage coverage of 11,313x vs. mean coverage of 4,539x ± s.d. 751x for contaminating microbial DNA). The putative prophage is located on the WG8 scaffold_29, bp 5,310-51,124. To identify putative capsid proteins beyond existing PFAM domains, we compiled major capsid proteins from the DJR lineage, recently highlighted as a group of non-tailed viruses both diverse and widespread in nature[47]. A custom HMM profile was created as in Kauffman *et al*. 2018, and predicted proteins from the putative prophage were compared to this profile using hmmsearch[64]. This search returned no hits, even with relaxed cutoffs of 30 on score. Consistently, a blastp search against the same database of DJR major capsid proteins returned no significant hits (big score <30 and *E*-value <1e^{-2}. *Halanaerobium* cells and viral particles were imaged at the Molecular and Cellular Imaging Center, Ohio State University, Wooster OH. An equal volume of 2x fixative (6% glutaraldehyde, 2% paraformaldehyde in 0.1 M potassium

phosphate buffer pH 7.2) was added directly to the culture post-induction. 30 µl of media was applied to a formovar and carbon coated copper grid for 5 minutes, blotted and then stained with 2% uranyl acetate for 1 minute. Samples were examined with a Hitachi H-7500 electron microscope and imaged with the SIA-L12C (16 megapixels) digital camera.

Code availability

All scripts and analyses necessary to perform metagenome assembly, EMIRGE, annotation and single-copy genes can be accessed from github (https://github.com/TheWrightonLab/metagenome_analyses).

Data availability

Halanaerobium isolate genome and MAGs are publicly available in the JGI Genome Portal database (http://img.jgi.doe.gov/) or in NCBI, see Supplementary Table 1 for accession numbers. All of the metagenomic nucleotide files used in this study are publicly available through JGI or NCBI, accession numbers are listed in Supplementary Data 1.

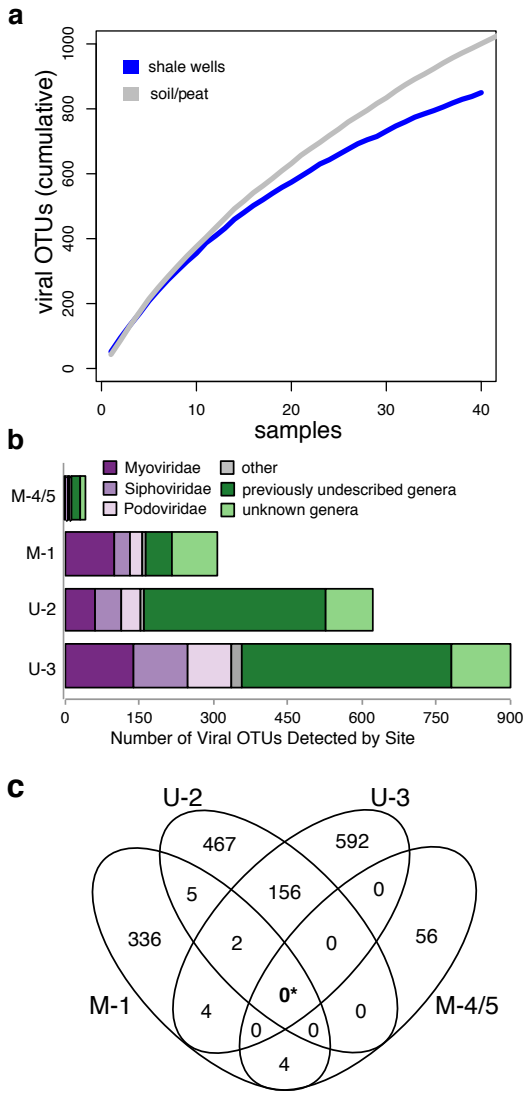

Fig. 1 Overview of viruses detected in hydraulically fractured shale wells. **a** Viral OTU accumulation curve, contrasting OTUs detected in a study of soil/peat samples[9] with viruses detected in hydraulically fractured shale wells. Only shale-derived viral sequences ≥ 10kb are compared to soil sequences. **b** Bar graphs show the number and taxonomy of viral OTUs detected at each site throughout the sampling periods. Data from wells M-4 and M-5 were combined for clarity. Viral populations that did not cluster with any reference genomes, but other viruses from hydraulically fractured wells, were defined as constituting 'previously undescribed genera.' Viral population singletons that did not cluster with any other viral populations were defined as 'unknown genera'. 'Other' includes viral populations with taxonomies: unclassified Caudovirales (n=24), dsDNA unclassified (n=4), dsDNA Corticoviridae (n=2), ssDNA Inoviridae (n=4), unclassified viruses (n=3).**c** Venn diagram showing the distribution of shared and site-specific viral OTUs. An asterix indicates that no viral OTUs are shared between all sites. (M-1, Marcellus-1, 1 well; M-4/5, Marcellus-4 and Marcellus-5, 2 wells, ;U-2, Utica-2, 1 well; U-3, Utica-3, 1 well).

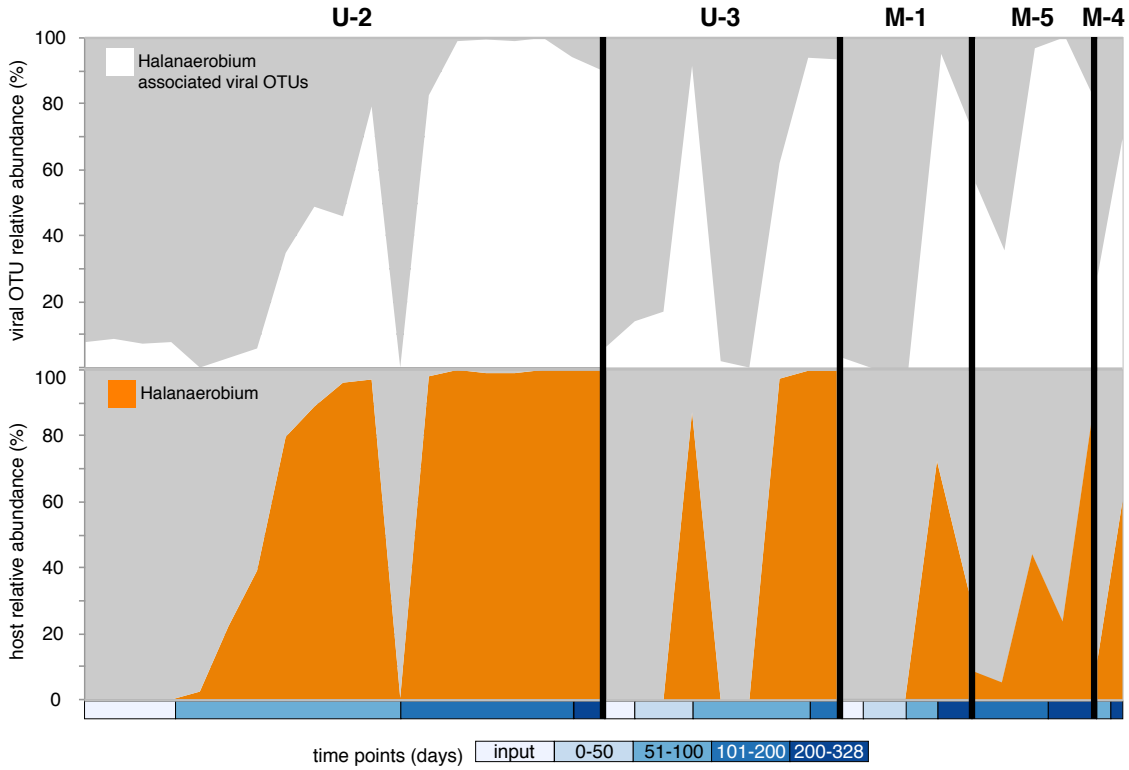

Fig. 2 Viral and *Halanaerobium* dynamics in hydraulically fractured wells. Top panel: Summed relative abundance of viral OTUs associated with *Halanaerobium* strains (in white) via CRISPR-Cas spacer links and hexamer-nucleotide frequency, temporally by site/well, shown as area. Bottom panel: Summed relative abundance of *Halanaerobium* hosts (in orange) inferred by 16S rRNA, temporally by site, shown as area. Viral OTU and *Halanaerobium* abundances are positively correlated (Spearman's rho = 0.828, $p<0.001$). Sampling time points denote days after hydraulic fracturing and are colored by the legend at bottom (U-2, n=19; U-3, n=8; M-1, n=5; M-4, n=7 samples).

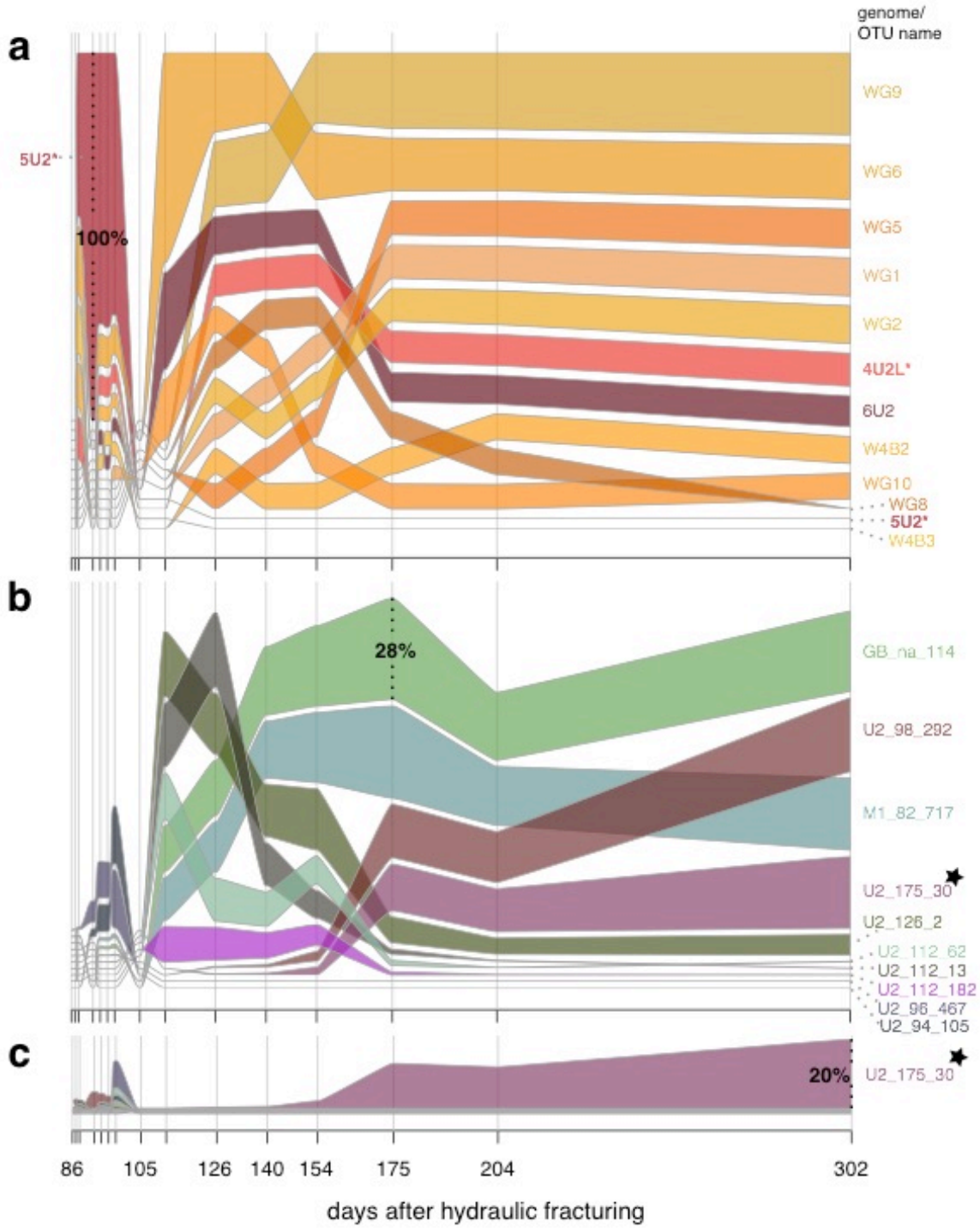

Fig. 3 Alluvial plots showing dynamics of *Halanaerobium* genomes (isolates and MAGs) and viruses within the Utica-2 well. Alluvia height (y-axis) is proportional to relative abundance; within each panel alluvia are ordered by rank at each time point. Genome and OTU names are listed to the right. A scale bar (%) marks the greatest relative abundance in each plot; the three plots are scaled identically. Alluvia marked with a star in panels b and c indicate the same viral OTU. **a** Dynamics of 12 *Halanaerobium* genomes (three MAGs, nine isolate genomes). Isolate genomes are in oranges, MAGs are in reds and

labeled with an asterix after the MAG name. **b** Dynamics of the 10 most abundant *Halanaerobium*-associated viruses based on CRISPR-Cas spacer matches and/or kmer frequency dissimilarity. **c** Dynamics of the 10 most abundant viruses linked to *Halanaerobium* in the Utica-2 well via CRISPR-Cas spacer matches.

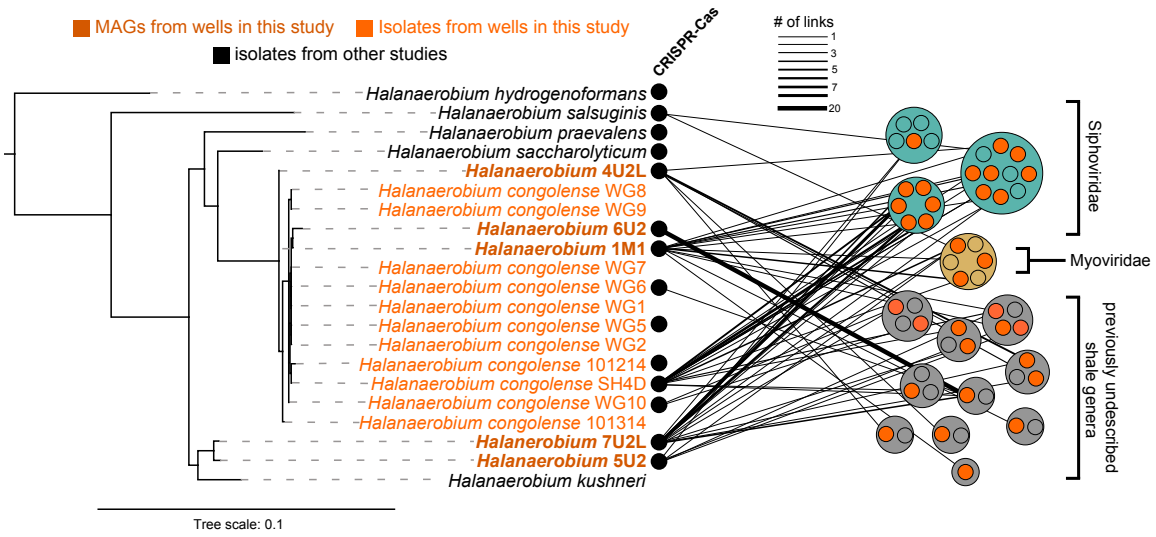

Fig. 4 Network of genomic links between viral genera and *Halanaerobium* hosts. Left, concatenated ribosomal protein tree for *Halanaerobium* isolate genomes and MAGs based on 16 ribosomal proteins chosen as single-copy phylogenetic marker genes. *Halanaerobium* hosts in orange are from hydraulically fractured wells from this study, *Halanaerobium* hosts in black are isolates from other environments/studies. Right, network of viral-host links based on CRISPR-Cas spacer matches. Each large circle represents a viral cluster approximating genus-level taxonomy. These include three genera within the Siphoviridae (teal), one genus within the Myoviridae (brown), nine previously undescribed genera and a 'singleton' genome (gray). Within these circles, each smaller circle represents a single viral OTU. Small circles in orange are viral OTUs genomically linked to *Halanaerobium*. Small circles without orange fill are viral OTUs within each genera without genomic links to *Halanaerobium*. Within a family, circles representing viral genera are randomly placed. Lines represent links between a Halanaerobium CRISPR-Cas spacer and a viral protospacer; line width denotes the number of links between a single virus and host; five or greater links between a *Halanaerobium* genome and a single virus is indicated with a red line. One spacer mismatch was allowed; 141 of the 151 spacer links were perfect matches.

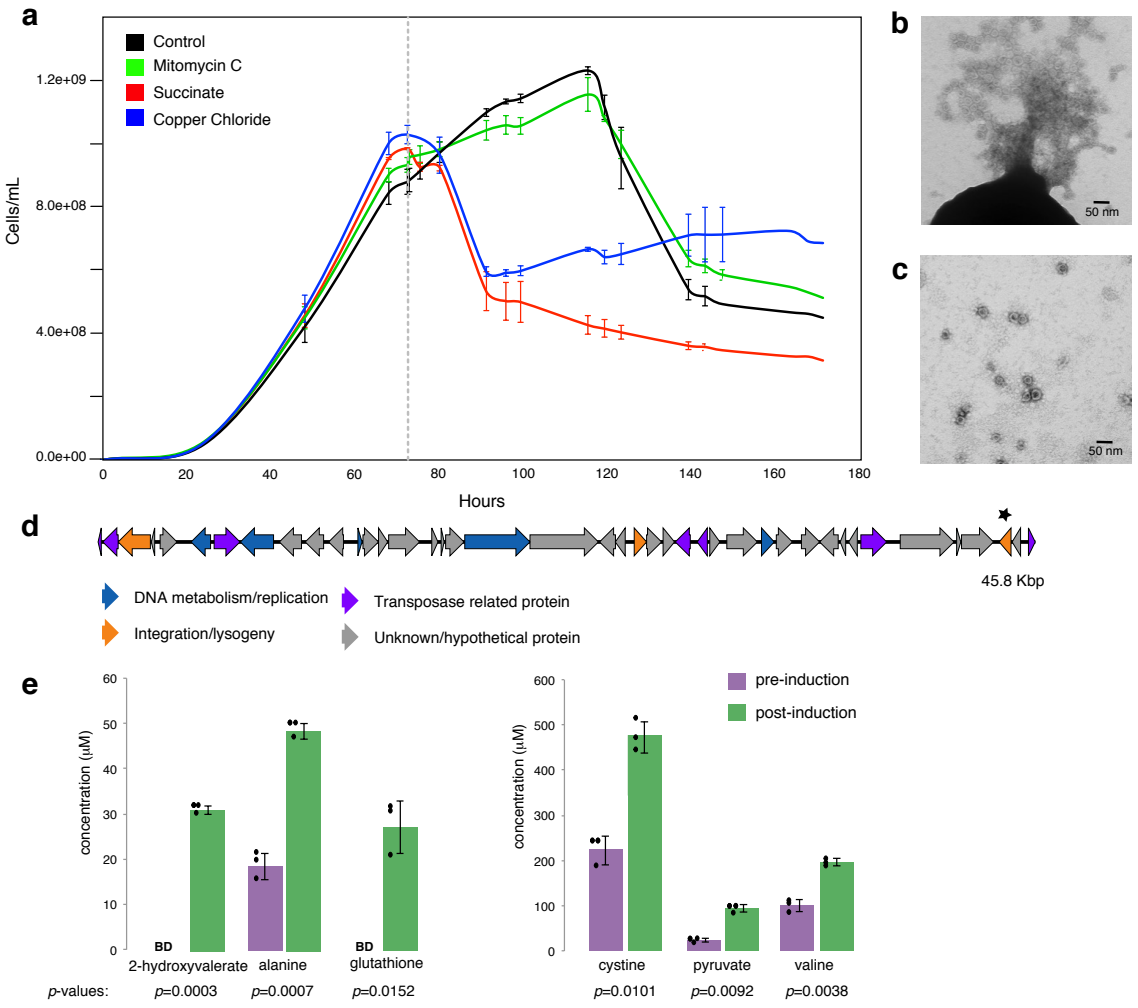

Fig. 5 *Halanaerobium* strain WG8 prophage induction. **a** Growth curve of *Halanaerobium* WG8 shown as cells per milliliter through time. Prophage induction was attempted using mitomycin C (green), succinate (red) and copper chloride (blue). Control (no chemical induction) is shown in black. The vertical, dotted grey line indicates the point of induction at 72 hours (n=3 independent samples for each treatment). Center values indicate the mean and error bars indicating one standard deviation. **b-c** TEM images post-succinate induction negatively stained with uranyl acetate. These experiments were repeated three times independently with similar results. **b** Lysis of a *Halanaerobium* WG8 cell, and release of viral-like particles (VLPs). Burst size was calculated as ~61 viruses per lytic event. **c** Individual VLPs are tailless, with 27.5 nm ± 3.8 nm in diameter capsids (n=219 capsid diameter measurements reporting the mean ± one standard deviation). **d** Genome of induced prophage from *Halanerobium* WG8. Genes related to DNA metabolism and replication are shown in blue, genes related to integration and lysogeny are shown in orange, transposase related genes in purple, and unknown or hypothetical genes are shown in grey. The star indicates the location of a superinfection exclusion gene. **e** Barchart of metabolites with significant increases in concentration after succinate induction (n=3 independent samples) (students paired two-sided t-test with unequal variance, showing the mean and error bars indicating one

standard deviation, black dots show the distribution of data, *p*-values for each comparison are below bars). Pre-induction concentrations are shown in purple; post-induction concentrations are shown in green. BD = below detection.

References

1 Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361, (2005).

2 Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578-6583, (1998).

3 De Smet, J. *et al.* High coverage metabolomics analysis reveals phage-specific alterations to *Pseudomonas aeruginosa* physiology during infection. *ISME J* **10**, 1823-1835, (2016).

4 Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews* **64**, 69-114, (2000).

5 De Smet, J., Hendrix, H., Blasdel, B. G., Danis-Wlodarczyk, K. & Lavigne, R. *Pseudomonas* predators: understanding and exploiting phage–host interactions. *Nat. Rev. Microbiol.*, 1-14, (2017).

6 Feiner, R. *et al.* A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol.* **13**, 641-650, (2015).

7 Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498-1261498, (2015).

8 Danovaro, R. *et al.* Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* **454**, 1084-1087, (2008).

9 Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, (2018).

10 Nigro, O. D. *et al.* Viruses in the oceanic basement. *mBio* **8**, (2017).

11 Pan, D. *et al.* Correlation between viral production and carbon mineralization under nitrate-reducing conditions in aquifer sediment. *ISME J* **8**, 1691-1703, (2014).

12 Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334-338, (2010).

13 Anderson, R. E., Brazelton, W. J. & Baross, J. A. Is the genetic landscape of the deep subsurface biosphere affected by viruses? *Front Microbiol* **2**, 219, (2011).

14 Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nature Microbiology*, 16146, (2016).

15 Booker, A. E. *et al.* Sulfide generation by dominant *Halanaerobium* microorganisms in hydraulically fractured shales. *mSphere* **2**, (2017).

16 Liang, R. *et al.* Metabolic Capability of a predominant *Halanaerobium* sp. in hydraulically fractured gas wells and its implication in pipeline corrosion. *Front Microbiol* **7**, 116, (2016).

17 Lipus, D. *et al.* Predominance and metabolic potential of *Halanaerobium* in produced water from hydraulically fractured Marcellus shale wells. *Applied and Environmental Microbiology*, AEM.02659-02616, (2017).

18 Mouser, P. J., Borton, M., Darrah, T. H., Hartsock, A. & Wrighton, K. C. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol Ecol* **92**, (2016).

19 Bhupathiraju, V. K., McInerney, M. J., Woese, C. R. & Tanner, R. S. *Haloanaerobium kushneri* sp nov., an obligately halophilic, anaerobic bacterium from an oil brine. *Int. J. Syst. Evol. Microbiol.* **49**, 953-960, (1999).

20 Bhupathiraju, V. K. *et al. Haloanaerobium salsugo* Sp-Nov, a moderately halophilic, anaerobic bacterium from a subterranean brine. *Int. J. Syst. Evol. Microbiol.* **44**, 565-572, (1994).
21 Brown, S. D. *et al.* Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium *Halanaerobium hydrogeniformans*. *Journal of Bacteriology* **193**, 3682-3683, (2011).
22 Kivisto, A. *et al.* Genome sequence of *Halanaerobium saccharolyticum* subsp. saccharolyticum strain DSM 6643T, a halophilic hydrogen-producing bacterium. *Genome Announcements* **1**, e00187-00113-e00187-00113, (2013).
23 Zeikus, J. G., Hegge, P. W., Thompson, T. E., Phelps, T. J. & Langworthy, T. A. Isolation and description of *Haloanaerobium praevalens* gen. nov. and sp. nov., an obligately anaerobic halophile common to Great Salt Lake sediments. *Curr Microbiol* **9**, 225-233, (1983).
24 Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, (2015).
25 Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics*, 1-13, (2016).
26 Leff, J. W. *et al.* Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl. Acad. Sci. U.S.A.*, (2015).
27 Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243, (2017).
28 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 1-20, (2016).
29 Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425-430, (2016).
30 Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* **46**, D708-D717, (2018).
31 Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K. & Mouser, P. J. Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus shale gas wells. *Environ. Sci. Technol.* **48**, 6508-6517, (2014).
32 Davis, J. P., Struchtemeyer, C. G. & Elshahed, M. S. Bacterial communities associated with production facilities of two newly drilled thermogenic natural Gas wells in the Barnett Shale (Texas, USA). *Microb Ecol* **64**, 942-954, (2012).
33 Murali Mohan, A., *et al.* Microbial community changes in hydraulic fracturing fluids and produced water from shale gas extraction. *Environ Sci Technol*, **47**, 13141-13150, (2013).
34 Wuchter, C., Banning, E. & Mincer, T. J. Microbial diversity and methanogenic activity of Antrim Shale formation waters from recently fractured wells. *Frontiers in Microbiol*, **4** (2013).
35 Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research* **45**, 39-53, (2017).

36 Wang, K., Wommack, K. E. & Chen, F. Abundance and distribution of *Synechococcus* spp. and cyanophages in the Chesapeake Bay. *Applied and Environmental Microbiology* **77**, 7459-7468, (2011).
37 Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Research* **23**, 111-120, (2013).
38 Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Applied and Environmental Microbiology* **59**, 3393-3399, (1993).
39 Stern, A. & Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *Bioessays* **33**, 43-51, (2010).
40 Howard-Varona, C. *et al.* Multiple mechanisms drive phage infection efficiency in nearly identical hosts. *ISME J*, (2018).
41 Sun, C. L., Thomas, B. C., Barrangou, R. & Banfield, J. F. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. **10**, 858-870, (2015).
42 Emerson, J. B. *et al.* Virus-host and CRISPR Dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* **2013**, 1-12, (2013).
43 Achigar, R., Magadán, A. H., Tremblay, D. M., Pianzzola, M. J. & Moineau, S. Phage-host interactions in *Streptococcus thermophilus*: Genome analysis of phages isolated in Uruguay and ectopic spacer acquisition in CRISPR array. *Sci. Rep.*, 1-9, (2017).
44 Gómez, P. & Buckling, A. Bacteria-phage antagonistic coevolution in soil. *Science* **332**, 106-109, (2011).
45 Levin, B. R. Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet*, (2010).
46 Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J* **7**, 1738-1751, (2013).
47 Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, 1-25, (2018).
48 Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through diverse mechanisms. *ISME J* **10**, 2854-2866, (2016).
49 Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 201800155, (2018).
50 Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J* **9**, 1352-1364, (2015).
51 John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports* **3**, 195-202, (2010).
52 Lever, M. A. *et al.* A modular method for the extraction of DNA and RNA, and the separation of DNA pools from diverse environmental sample types. *Front Microbiol* **6**, 1281, (2015).

53 Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033-1034, (2012).
54 Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W. & Banfield, J. F. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology* **12**, R44, (2011).
55 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357-359, (2012).
56 Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology*, (In Review).
57 Leplae, R., Hebrant, A., Wodak, S. J. & Toussaint, A. ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Research* **32**, D45-49, (2004).
58 Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498-2504, (2003).
59 Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009-1010, (2011).
60 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature Publishing Group* **13**, 722-736, (2015).
61 Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.* **78**, 4430-4442, (2006).
62 Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**, 470-483, (2009).
63 Williamson, S. J., Houchin, L. A., McDaniel, L. & Paul, J. H. Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Applied and Environmental Microbiology* **68**, 4307-4314, (2002).
64 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, (2011).

Correspondence and requests for materials should be addressed to M.J.W.

Acknowledgements
R.A.D., M.A.B, D.M.M, A.E.B, A.J.H, P.J.M., K.C.W. and M.J.W. are partially supported by funding from the National Sciences Foundation Dimensions of Biodiversity (award no. 1342701). R.A.D., M.A.B., D.M.M., A.E.B., K.C.W. and M.J.W. also received support from Dow Microbial Control for this work. Samples from wells M4 and M5 were provided by the Marcellus Shale Energy and Environment Laboratory (MSEEL) funded by Department of Energy's National Energy Technology laboratory (DOE-NETL) grant no. DE-FE0024297. Metagenomic sequencing for this research was performed by the DOE Joint Genome Institute (JGI) via a large-scale sequencing award to K.C.W (Award no. 1931). Metabolite support was provided by Environmental Molecular Sciences Laboratory (EMSL) support via a JGI–EMSL Collaborative Science Initiative awarded to K.C.W (Award no. 48483) and an EMSL instrument time award to M.J.W. (Award no. 49615). Both JGI and EMSL facilities are sponsored by the Office of Biological and Environmental Research and operated under contracts nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). M.B.S. was partially supported by a Gordon and Betty Moore Foundation grant (no. 3790).

Author contributions
R.A.D., K.C.W. and M.J.W. designed the study. A.J.H. and P.J.M. collected the samples. R.A.D., R.A.W. and M.A.B. performed bioinformatic analyses. D.M.M., A.E.B. and M.D.J. conducted laboratory induction analyses, while D.W.H. performed quantitative metabolite NMR measurements. T.M. conducted electron microscopy on *Halanaerobium* cultures. J.D.M. and K.W. participated in constructive manuscript discussions that resulted in an improved manuscript. M.J.W., K.C.W., M.B.S., S.R. and R.A.D. integrated the data and drafted the manuscript. All authors reviewed the results and approved the manuscript.

Competing interests
The authors declare no competing interests.