Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales

Rebecca A. Daly[1], Mikayla A. Borton[1], Michael J. Wilkins[1,2], David W. Hoyt[3], Duncan J. Kountz[1], Richard A. Wolfe[1], Susan A. Welch[2], Daniel N. Marcus[1], Ryan V. Trexler[4], Jean D. MacRae[5], Joseph A. Krzycki[1], David R. Cole[2], Paula J. Mouser[4] and Kelly C. Wrighton[1]*

Hydraulic fracturing is the industry standard for extracting hydrocarbons from shale formations. Attention has been paid to the economic benefits and environmental impacts of this process, yet the biogeochemical changes induced in the deep subsurface are poorly understood. Recent single-gene investigations revealed that halotolerant microbial communities were enriched after hydraulic fracturing. Here, the reconstruction of 31 unique genomes coupled to metabolite data from the Marcellus and Utica shales revealed that many of the persisting organisms play roles in methylamine cycling, ultimately supporting methanogenesis in the deep biosphere. Fermentation of injected chemical additives also sustains long-term microbial persistence, while thiosulfate reduction could produce sulfide, contributing to reservoir souring and infrastructure corrosion. Extensive links between viruses and microbial hosts demonstrate active viral predation, which may contribute to the release of labile cellular constituents into the extracellular environment. Our analyses show that hydraulic fracturing provides the organismal and chemical inputs for colonization and persistence in the deep terrestrial subsurface.

Shale gas accounts for one-third of natural gas energy resources worldwide. It has been estimated that shale gas will provide half of the natural gas in the USA, annually, by 2040, with the Marcellus shale in the Appalachian Basin projected to produce three times more than any other formation[1]. Recovery of these hydrocarbons is dependent on hydraulic fracturing technologies, where the high-pressure injection of water and chemical additives generates extensive fractures in the shale matrix. Hydrocarbons trapped in tiny pore spaces are subsequently released and collected at the wellpad surface, together with a portion of the injected fluids that have reacted with the shale formation. The mixture of injected fluids and hydrocarbons collected is referred to as 'produced fluids'.

Microbial metabolism and growth in hydrocarbon reservoirs has both positive and negative impacts on energy recovery. Whereas stimulation of methanogens in coal beds enhances energy recovery[2], bacterial hydrogen sulfide production ('reservoir souring') decreases profits and contributes to corrosion and the risk of environmental contamination[3]. Additionally, biomass accumulation within newly generated fractures may reduce their permeability, decreasing natural gas recovery. Despite these potential microbial impacts, little is known about the function and activity of microorganisms in hydraulically fractured shale.

Initial work by our group and others[4–9] used single marker gene analyses to identify microorganisms from several geographically distinct shale formations. These analyses showed similar halotolerant taxa in produced fluids several months after hydraulic fracturing. To assign functional roles to these organisms, we conducted metagenomic and metabolite analyses on input and produced fluids up to a year after hydraulic fracturing (HF) from two Appalachian basin shales, the Marcellus and Utica/Point Pleasant (Utica) formations. Although an earlier metagenomic study examined shale-produced fluids[10], the microbial communities were only sampled for nine days after HF. Here, we have reconstructed the first genomes from fractured shale, examining the microbial metabolisms sustained in these engineered, deep subsurface habitats over a period of 328 days. We provide evidence for metabolic interdependencies, and describe chemical and viral factors that control life in these economically important ecosystems. Our results show microbial degradation of chemical additives, the potential for microbially induced corrosion and the formation of biogenic methane, all of which have implications for the sustainability of energy extraction.

Reconstruction of persisting shale genomes

Our earlier study surveyed microbial community structure in fluids from three hydraulically fractured Marcellus shale wells[5]. Five fluid samples from a single well were chosen for paired metagenomic and metabolite analyses, as these samples represented three phases of the energy extraction process. Fluid samples were collected from input materials, and at early (7 and 13 days) and late (82 and 328 day) time points (Fig. 1a) following HF. Microbial community changes during these phases corresponded to increasing salinity (Fig. 2 and Supplementary Table 1). From these Marcellus shale fluid samples we recovered 34 genomic bins, composed of 29 unique genomes (Fig. 1b and Supplementary Table 2). A high percentage of sequenced reads mapped to the assemblies (89–99%) (Supplementary Table 3), signifying that the underlying data were well represented. We also validated that the assembled genomes

[1]Department of Microbiology, The Ohio State University, Columbus, Ohio 43214, USA. [2]School of Earth Sciences, The Ohio State University, Columbus, Ohio 43214, USA. [3]EMSL, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. [4]Department of Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, Ohio 43214, USA. [5]Department of Civil and Environmental Engineering, University of Maine, Orono, Maine 04469, USA. *e-mail: wrighton.1@osu.edu

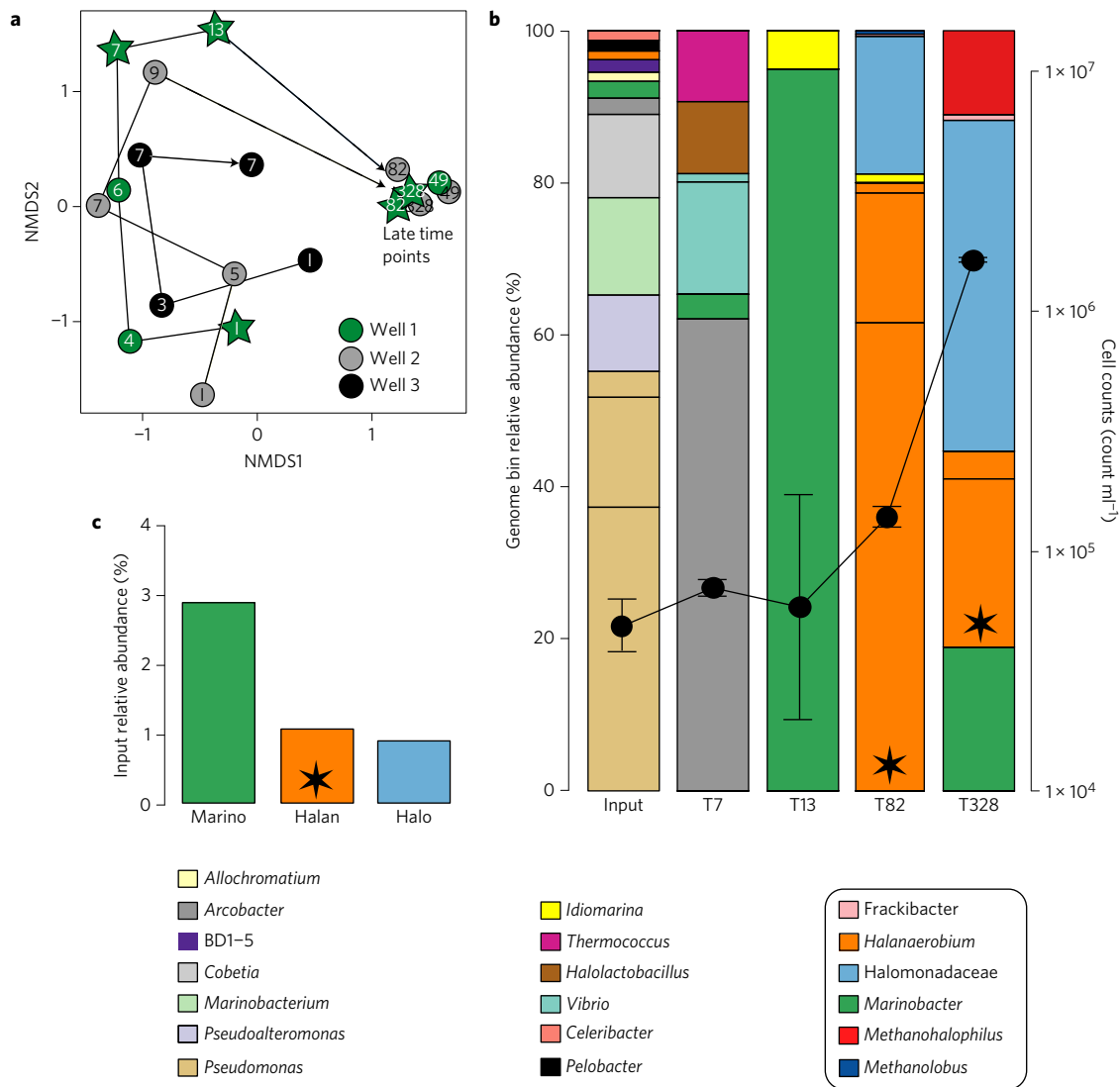**Figure 1 | Genomic sampling from hydraulically fractured Marcellus shale fluids over time. a**, Non-metric multidimensional scaling ordination of 16S rRNA gene data from our previous study[5] show similar community trajectories across wells. The Marcellus fluid samples analysed here for metagenomic and metabolite analyses are indicated by green stars. **b**, Genome bin relative abundance at each time point, with taxa coloured according to the legend and persisting members outlined in a black box. For each time point, cell count data are overlaid with error bars representing the mean ± s.d. of technical replicates ($n = 3$). **c**, 16S rRNA gene relative abundance of persisting taxa present at low abundance in the input fluids. The star in the *Halanaerobium* orange bars (**b**,**c**) denotes an identical genome recovered from input, day 82 and day 328 samples.

reflected the microbial identities and abundances in the unassembled reads, by comparing genome bin relative abundance to reconstructed near-full-length 16S rRNA genes[11] (Supplementary Data File 1).

Consistent with our earlier taxonomic study, six halotolerant bacterial and archaeal members became enriched at later time points (82 and 328 days). We recovered six *Halanaerobium*, two Halomonadaceae, four *Marinobacter*, one *Methanohalophilus*, one *Methanolobus* and two bins from Halobacteroidaceae (Supplementary Table 2 and Supplementary Data File 2). Each of these taxa contain halotolerant and thermotolerant members. Environmental sequences closely related to 16S rRNA genes recovered here were similar to those also recovered from other hydrocarbon reservoirs or hypersaline environments (Supplementary Data File 3). For each of these persisting taxa (Fig. 1b) we obtained a representative genome that was at least 90% complete and, with the exception of *Marinobacter* (which contained several low-abundance *Marinobacter* strains), had less than 1% estimated contamination (Supplementary Table 4). The

Halobacteroidaceae genus lacked closely related 16S rRNA genes (~94% identity) and genomes (76% average nucleotide identity, ANI), suggesting this organism may be unique to shales (Supplementary Fig. 1). Following the naming convention for near-complete (>95% sampling) genomes from metagenomic data sets[12], we propose the genus name *Candidatus* Frackibacter based on the colloquial name for hydraulic fracturing, or 'fracking'. We infer that changes in membership at these later time points were due to growth of these specific taxa rather than DNA persistence in the environment, as cellular biomass increased in this well (Fig. 1b).

Emphasizing the persistence of specific taxa during energy extraction, members of the terminal community were identified from the input fluid through either identical genomes or closely related 16S rRNA genes (Fig. 1c). For instance, a *Halanaerobium* genome detected at both days 82 and 328 had an identical genome in the input fluid (ANI ~99% with gene synteny) (Supplementary Table 5). Conversely, we did not detect lower-abundance members of the terminal community (for example, methanogens and *Candidatus* Frackibacter, <2% 16S rRNA gene
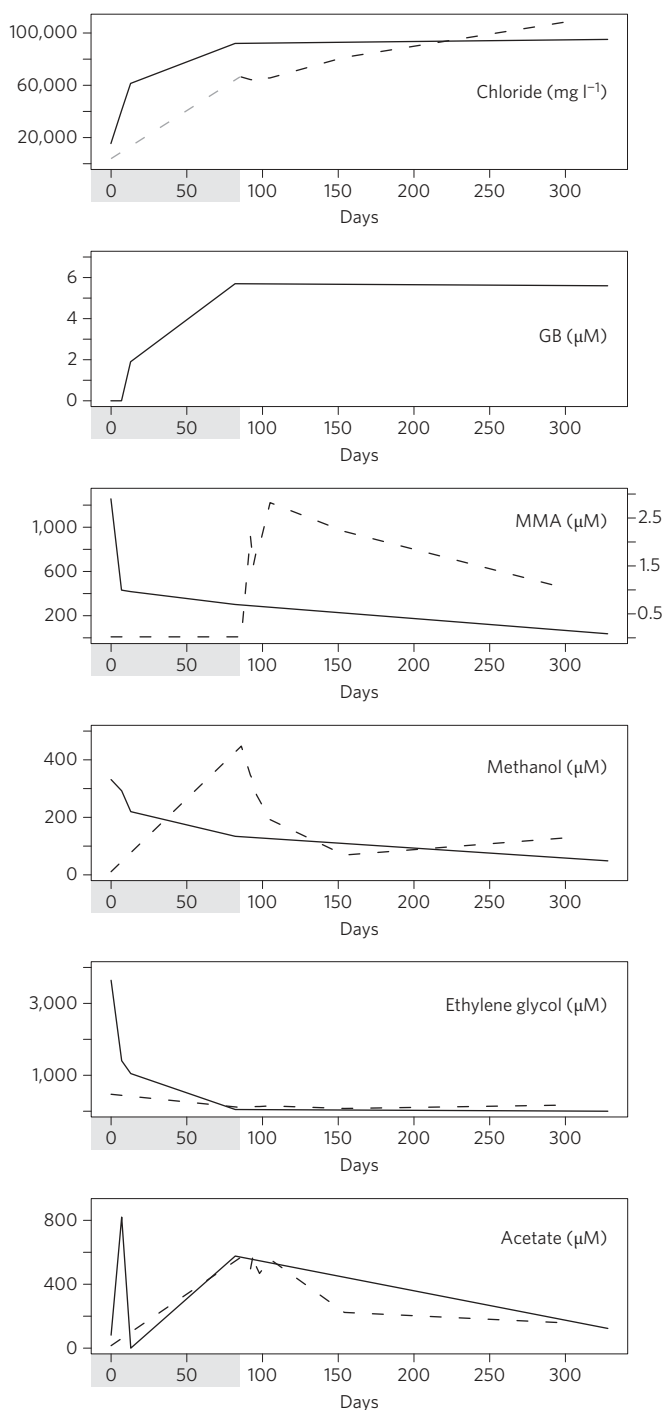
Figure 2 | Quantification of metabolites identified by ¹H NMR and chloride in fluids from Marcellus and Utica hydraulically fractured shales (for additional data see Supplementary Table 1). Time 0 on the x axis denotes input HF fluids, with the time the Utica well was shut in shown as grey shading. Marcellus fluids are denoted by solid lines and Utica fluids by dashed lines. For the Utica, initial chloride values were estimated based on the freshwater source (grey dashed line). MMA concentrations are shown with two axes, with the left axis for Marcellus fluids (solid lines) and the right axis for Utica fluids (dashed lines).

abundance) in the input fluids, probably because they were below our detection limit (Supplementary Data File 1b). This finding is the first to demonstrate that HF creates a habitat where low-abundance microorganisms are injected into the deep subsurface, bloom, and persist despite biocide addition[13], elevated

temperatures (65 °C) and pressures (at least 25 MPa), and salinities that ultimately become briny.

Glycine betaine and chemical additives fuel methanogenesis

Halite dissolution from the shale matrix drives large salinity increases in the produced fluids[9], so organisms must have adaptations for tolerating a broad salinity range (Fig. 2). We recovered multiple osmoprotectant strategies from all genomes (Supplementary Table 6 and Supplementary Discussion). Our metabolite data show that, of the known osmoprotectants[14], glycine betaine (GB) was present in the fluids, but mannitol, sorbital, ecotine and trehalose were not detected (Supplementary Table 1). Consistent as a response to salinity, GB was below detection in the input and early Marcellus shale fluids, but reached a maximum concentration at day 82 that was maintained at day 328 (Fig. 2).

Uptake and *de novo* synthesis of GB were features encoded in all near-complete genomes recovered over the last two time points. GB synthesis is encoded in the *Methanohalophilus* genome by a glycine pathway (via sarcosine and dimethylglycine intermediates) and in the Halomonadaceae and *Marinobacter* genomes by a pathway from choline (Supplementary Discussion). Choline, a common chemical additive in fracturing fluids, was exogenously provided in the input fluids and consumed by day 13 when *Marinobacter* became abundant (Fig. 1b and Supplementary Table 1). Collectively, our paired metagenomic and metabolite findings show the production and uptake of GB is a halotolerance mechanism widely used by organisms in fractured shales.

Microbially synthesized GB, available extracellularly in the fluids, may be degraded by both of the recovered obligate fermenters (*Halanaerobium*, *Candidatus* Frackibacter) (Supplementary Discussion). *Candidatus* Frackibacter has two mechanisms for reducing GB. The first demethylates GB and oxidizes the methyl group via the Wood–Ljungdahl pathway, producing trimethylamine (TMA; Supplementary Fig. 2)[15]. The second pathway, also present in some shale *Halananerobium* genomes, uses a GB reductase (*grdHI*), producing TMA and acetate via a Stickland reaction[16] (Supplementary Fig. 3). Notably, GB reduction is not widely encoded in isolated *Halanaerobium* genomes, being present in only 36% of the published genomes (5 of 14, http://img.jgi.doe.gov, April 2016). In the Marcellus shale, a persisting *Halanaerobium* (*Halan*-2; T82, T328) genome is the only one of our three capable of GB reduction. GB fermentation using microbially produced metabolites, rather than a dependency on input fluid chemicals, may sustain life in shales long after hydraulic fracturing.

GB fermentation yields TMA, which we infer is rapidly consumed by methanogens present at the last two time points. Each of the recovered shale methanogen genomes (*Methanolobus* and *Methanohalophilus*) has pathways for utilizing TMA, dimethylamine (DMA), monomethylamine (MMA) and methanol, but cannot use GB, hydrogen/carbon dioxide or acetate (Supplementary Discussion). In addition to microbial synthesis, HF input fluids also contain high concentrations of methylotrophic substrates (1.2 mM MMA and 331 µM methanol) that could support methanogenesis (Fig. 2). It is possible that these compounds are also assimilated as a nitrogen source (MMA) or are oxidized by *Pseudomonas* and *Marinobacter* (methanol) at earlier time points[17]. Although *Methanohalophilus* 16S rRNA genes have been reported from Antrim[8,9] and Burket/Geneseo[4] fractured shales, our genomic and metabolite findings identify the endogenous and exogenous sources of methyltrophic substrates, show their co-occurrence with methanogens, and confirm the metabolic pathways for methanogenesis.

Metabolisms impacting energy extraction

In addition to containing substrates that could support biogenic methane production, HF input fluids contain high concentrations
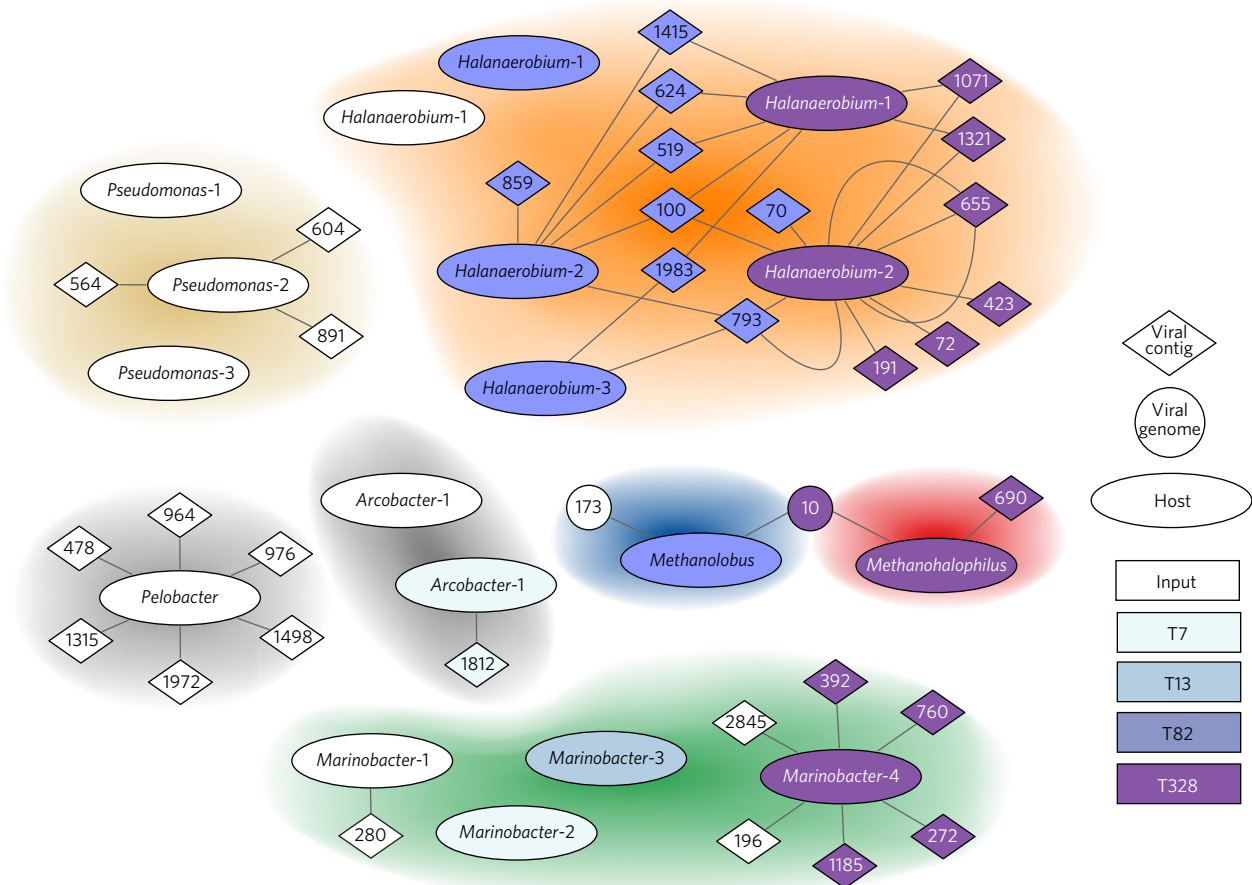
Figure 3 | Genomic links between viruses and their microbial hosts. Network of viral–host CRISPR links with host genomes represented by ovals, viral contigs by diamonds and closed viral genomes by circles. Colours represent sampling times. Clouds of colour indicate host taxonomy (Fig. 1b). Each edge represents a perfect match between a host CRISPR loci spacer sequence and a viral contig. Numbers in diamonds and circles denote the contig numbers.

of organic substrates such as sucrose (0.3 mM) and ethylene glycol (3.6 mM) (Fig. 2 and Supplementary Table 1). The capacity to respire sucrose is widely encoded in our genomes (for example, *Vibrio*, *Pseudomonas* and *Marinobacter*), consistent with its consumption by day 7. Ethylene glycol is consumed over time, and is not detected at the last time point. This substrate is probably aerobically oxidized by *Marinobacter* and *Vibrio* at the early time points[18] (alcohol dehydrogenase), and fermented by *Halanaerobium* (propanediol dehydratase, acetalaldehyde dehydrogenase) later to yield ethanol, hydrogen and acetate. *Candidatus* Frackibacter also has the capacity to produce acetate via GB fermentation, homoacetogenesis (H_2/CO_2) and sugar fermentation. Consistent with the possibility for GB, sugar and ethylene glycol fermentation at later time points, ethanol and acetate increased at day 82, when *Halanaerobium*, *Candidatus* Frackibacter and methanogens were co-enriched.

Halanaerobium are the dominant members in produced fluids from Barnett, Marcellus, Burket/Geneseo and Antrim fractured shales[4,6,7,9]. Our shale-hosted *Halanaerobium* genomes also have the capacity to ferment amino acids (for example, alanine; Supplementary Discussion), sucrose, fructose, glucose and maltose (Supplementary Table 7). Biofilm formation may be an important adaptation enabling the dominance of *Halanaerobium* across hydraulically fractured shales. Although biofilm-related genes are not detected in all surface *Halanaerobium* genomes[19], these shale genomes encode genes for flagellar motility and cellular aggregation (for example, polysaccharide production and diguanylate cyclase)[20] (Supplementary Table 7).

Biofilm, organic acid and H_2 production, together with the capacity to reduce thiosulfate to sulfide (using three copies of the rhodanese-like thiosulfate:cyanide sulfur-transferase gene[3]), implicate a role for shale *Halanaerobium* in steel corrosion and reservoir souring[21]. Additionally, the near-complete Halomonadaceae genome also encodes multiple thiosulfate sulfur-transferase genes, which while not previously reported in these taxa, are implicated in thiosulfate disproportionation, producing sulfide and sulfite[22]. Current microbial corrosion diagnostic practices often rely on detecting the presence of dissimilatory sulfate-reducing genes or measuring sulfate-reducing metabolic potential. However, we did not identify any sulfate-reducing genes in the Marcellus data set, suggesting the need to include alternative biological mechanisms of sulfide production for characterizing microbial corrosion potential.

Owing to the economic importance of hydrocarbons, we analysed shale-produced fluids for degradation pathways, and confirmed the presence of benzene, toluene, ethylbenzene, xylene and naphthalene (BTEX-N) in all fluid samples, while decane was detected in all but the input[5]. We failed to recover any genes for anaerobic hydrocarbon degradation, but the near-complete *Marinobacter*, Halomonadaceae, *Pseudomonas* and *Vibrio* genomes (input or early samples) had the capacity for aerobic hydrocarbon oxidation (Supplementary Discussion).

Of the persisting members, the *Marinobacter* genome encodes the capacity for alkane and BTEX-N degradation, whereas the Halomonadaceae genome lacks the first steps of these pathways but has genes for subsequent degradation (Supplementary Fig. 4). These two taxa often co-occur in saline hydrocarbon habitats[23],

including Barnett[6] and Marcellus[7] shale-produced fluids. Of our recovered genomes, these two encode the greatest metabolic versatility, enabling the use of a wide range of carbon sources (for example, acetate, lactate and hexose sugars) with O_2 or nitrate as possible electron acceptors. However, given the lack of detectable nitrate (Supplementary Table 1), we postulate that these facultative anaerobes utilize fermentative metabolisms once dissolved O_2 associated with HF has been depleted[24,25].

Active viral predation in the deep subsurface

Our results indicate that viral-mediated cell lysis is a mechanism to explain how an intracellular osmoprotectant, like GB, was detected extracellularly in fluids. We recovered 331 viral contigs including 21 closed, circular viral genomes (Supplementary Fig. 5). A comparison of contigs across samples showed that 318 viral contigs were unique, with only 13 viral contigs shared across time points. Of the viral contigs, 86% belonged to members of the Caudovirales, tailed dsDNA viruses, with Myoviridae (44%) and Siphoviridae (26%) families predominating. Previously, only viral reads and prophage genome fragments have been reported[10,26,27].

We mined our microbial genomes for the presence of CRISPR-Cas systems (clustered regularly interspaced short palindromic repeat-CRISPR associated), which act as an acquired immunity to viruses[28] (Supplementary Table 8). CRISPR-Cas frequency estimates range from 81% of archaeal and 40% of bacterial genomes in cultivated microbes[29], to 10% of archaeal and bacterial genomes in metagenomic data sets[30]. In contrast, 100% of the three archaeal and 84% of the 31 bacterial genome bins of the Marcellus samples had evidence of a CRISPR-Cas system, with type 1 being the most prevalent (Supplementary Table 8). In fact, all microbial genomes at the last time point had a CRISPR-Cas system, signifying that viral immunity may be an important adaption for persistence in hydraulically fractured shales.

Comparing CRISPR arrays in microbial hosts to viral contig sequences allowed us to reconstruct a history of viral encounters, and link 34 viral contigs to 11 microbial genomes (Fig. 3 and Supplementary Fig. 5). Before our findings, the greatest number of reported CRISPR links within a data set was five, from a three-year study in a hypersaline lake[31]. Our data showed that viral host specificity varied, with viruses linked to multiple species within a genus (for example, Halanaerobium, Marinobacter) and a single viral genome linked to two methanogen genera (Fig. 3). We observed an increase in the number of CRISPR spacers within two Halanaerobium genomes between days 82 and 328, demonstrating that adaptive viral resistance probably occurred during this time span (Supplementary Fig. 6 and Supplementary Table 8). Our metagenomic data demonstrate that viral predation and host-acquired immunity are active processes in the deep terrestrial subsurface.

Strain metabolic diversity across shales

We examined fluid metabolites collected over 302 days after HF from the Utica shale, a geographically and stratigraphically distinct Appalachian Basin formation. Despite these differences, metabolite trends in the Marcellus and Utica produced fluids were similar. For instance, methanol and ethylene glycol were detected in input fluids and salinity increased over time in both shales (Fig. 2). Unlike the Marcellus, MMA was not detected in the input but was produced over time in the Utica produced fluids, suggestive of ongoing GB production, fermentation and subsequent methanogenesis. However, due to the chemical complexity of the Utica produced fluids, we could not confirm the presence of GB.

To validate that fractured shales harbour microorganisms that produce methane from GB fermentation products, we amended Utica produced fluids with GB. The produced fluid sample was collected 96 days after HF, comparable to our Marcellus sample, where the co-occurrence of GB fermenting and methanogenic organisms

was first detected (day 82). Relative to the produced fluids, the addition of GB enriched for Methanohalophilus (70%) and three Halanaerobium genomes (~21, 3 and 0.5%) (Supplementary Fig. 7). The presence of a GB reductase system probably explains the changes in relative abundance within these Halanaerobium, as the dominant genome in the produced fluids lacked grdI (decreasing from 51 to 3%). This finding demonstrates the power of genome-centric metagenomics to partition local microdiversity, explaining the co-occurrence of strains with distinct functional roles.

In the enrichment, GB fermentation produced TMA, DMA and MMA in low amounts, probably due to active consumption by Methanohalophilus (Supplementary Tables 9 and 10). Compared to the unamended control, GB addition produced 6.5 times more methane per day. Collectively, our Marcellus field and Utica laboratory data provide evidence that GB synthesis and subsequent fermentation supports biogenic CH_4 in hydraulically fractured shales.

Comparative genomics showed that the dominant Methanohalophilus and Halanaerobium near-complete genomes (Supplementary Table 2) in the Utica enrichment were closely related strains[32] to the Marcellus genomes (~99% ANI) (Supplementary Table 5 and Supplementary Data File 2). In contrast, ANI values between the Marcellus and Utica Methanohalophilus and other sequenced species (M. mahii and M. halophilus, both isolated from surface waters), were ~91 and 92%, respectively. Methanohalophilus CRISPR array comparisons identified a single spacer sequence shared between the Marcellus produced fluids and Utica GB enrichment genomes; the two other non-shale-derived Methanohalophilus genomes lacked CRISPR-Cas systems. Two viral contigs also had high sequence identity (>95%), showing that these shales share genetically similar viruses. Together, our data demonstrate that environmental filtering results in populations, metabolisms and viral processes shared between these two geographically distinct fractured shale ecosystems.

Conclusion

Resolving genomes from Marcellus and Utica produced fluids unveiled microbial metabolisms, adaptations and viral predation resistance mechanisms in fractured shales. From 16S rRNA gene analyses we could not have predicted the role Halanaerobium strains play in fermenting GB and HF chemical additives such as ethylene glycol, nor would we have associated the Halmonadaceae with detrimental sulfide production. Our genomic analyses show that closely related strains are niche differentiated. For instance, GB addition selected for the only Halanaerobium genome with GB reduction capacity. We identified the metabolic capabilities of Candidatus Frackibacter, unique to fractured shales, which can also ferment GB. Our metagenomic data revealed a possible role for viruses in the top-down (predation and lysis) and bottom-up (release of cellular contents; for example, GB) control of microbial communities in fractured shale. Notably, unlike earlier studies, all host genomes recovered at the last time point contained a CRISPR-Cas system. We also identified active host responses to viral predation, including new spacer incorporation over time. Together, our viral findings demonstrate the probable importance of CRISPR-Cas-mediated immunity for microbial persistence in fractured shales.

Here, we show that hydraulic fracturing provides the organisms, chemistry and physical space to support microbial ecosystems in ~2,500-m-deep shales (Fig. 4). Ultimately, our metagenomic and metabolite results indicate that adaptation to high salinity, metabolism in the absence of oxidized electron acceptors, and viral predation are potential controlling factors mediating long-term microbial metabolism during energy extraction from fractured shales. This study highlights the resilience of microbial life to adapt to, and colonize, a habitat structured by physical and chemical features very different from their origin.
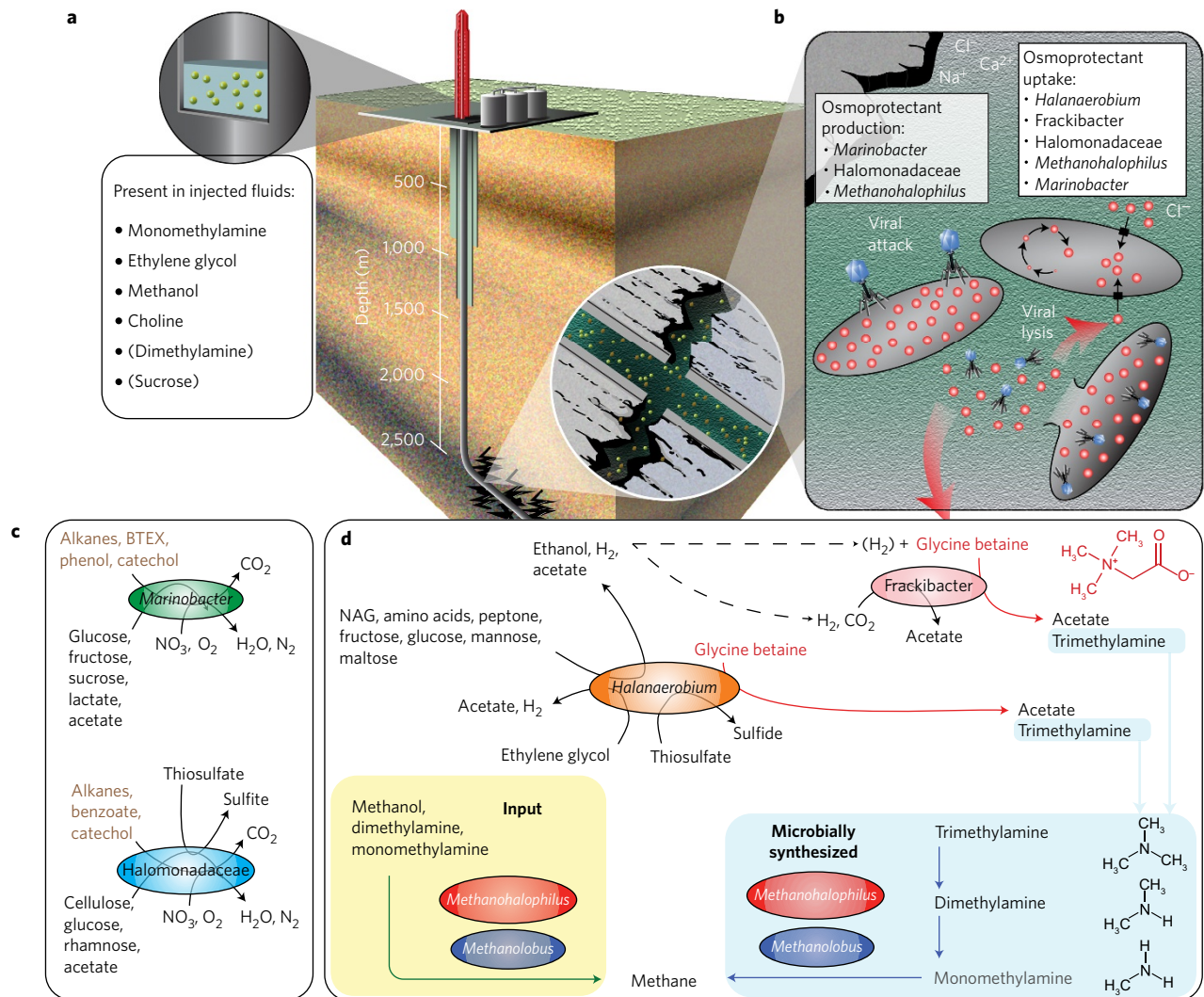
Figure 4 | Interconnected metabolisms catalysed by persisting microorganisms in hydraulically fractured shales. a, HF input fluids from both Marcellus and Utica shales contain substrates that sustain microbial metabolism. Parentheses indicate metabolites detected in one shale. **b,** Microorganisms in shales adapt to high salinities by producing and using osmoprotectants such as GB (red circles), which can be released into fluids by viral lysis. **c,** *Marinobacter* and Halomonadaceae have the potential to aerobically oxidize hydrocarbons and respire sugars using nitrate and oxygen as electron acceptors. **d,** *Candidatus* Frackibacter and *Halanaerobium* ferment GB, yielding trimethylamine, which supports methanogenesis by *Methanohalophilus* and *Methanolobus* (blue box). Methylamines and methanol in the input fluids can also support methanogenesis (yellow box).

Methods

Sample collection and fluid chemistry. Our earlier study characterized some of the geochemistry and conducted 16S rRNA gene surveys from fluids collected (June 2012 to May 2013) from three Marcellus gas wells located in Pennsylvania, USA[5]. Hydraulic fracturing (input, noted as T0) and shale-produced fluids were collected from well heads (days 3–14) and gas–fluid separators (49, 82 and 328 days), with fluids from well 1 used for more detailed metagenomics and NMR metabolite analyses here. For our Utica samples, injected fluids and produced fluids from gas–fluid separators[4,8,9] were collected between July 2014 and May 2015 from an oil–gas well in Ohio, USA. The gas–fluid separators at the Marcellus and Utica sites had a capacity of ∼5,560 l, approximately half gas and half produced fluids (2,780 l). Flow rates ranged from ∼380,000 l per day at early time points to ∼190,000 l per day at later time points (Marcellus day 328), with an estimated maximum residence time of 8 h on the days of sampling. Additionally, 16S rRNA sequences from key taxa we genomically sample here were either sampled here at earlier time points directly from the well head (for example, *Halomonas* and *Marinobacter*), or also recovered from other shale produced fluids where samples were collected exclusively from the well head (for example, *Halanaerobium*, *Methanolobus* and *Methanohalophilus*[9]).

Hydraulic fracturing included the injection of freshwater amended with chemicals, proppant and the addition of 20% recycled produced fluids for the Marcellus (not Utica). Notably, unlike the Marcellus wells, the Utica well was shut in for 86 days after fracturing, before initiating fluid and hydrocarbon collection at the surface (denoted on Fig. 2). Input and produced fluid samples (1 l) were collected in sterile bottles filled to capacity. As described elsewhere[5], ethoxylated surfactants and

hydrocarbons were assessed in Marcellus fluids using liquid chromatography quadrupole time of flight with electro-spray ionization (Agilent Technology) and gas chromotography (Hewlett Packard), respectively. Conductivity and pH were measured on unfiltered fluids in the field using Orion star probes, while fluid dissolved anions (F$^-$, acetate, formate, Cl$^-$, Br$^-$, NO$_3^-$, SO$_4^{2-}$, oxalate, S$_2$O$_3^{2-}$ and PO$_4^{3-}$) were analysed using a Thermoscientific Dionex ICS-2100 ion chromatograph with the exception of the Utica input fluid. Samples for ion chromatography were diluted by a factor of 10–100, due to the high salinity. The Utica input fluid had high viscosity, which precluded analysis by ion chromatography. Conductivity and Cl$^-$ were not measured for this input sample. NO$_3^-$ and SO$_4^{2-}$ in the Utica input were measured on unfiltered samples with a HACH DR/890 portable colorimeter using cadmium reduction method 8171 and turbidimetric method 8051, respectively. Fluid samples for major and trace cations (Na, Mg, K, Ca, Si, Sr, Ba, Li, Mn and Fe) were acidified immediately after filtration to ∼0.5% with nitric acid and then analysed using a Perkin Elmer Optima 4300DV inductively coupled plasma optical emission spectrometer. The charge balance discrepancy for the input samples is probably due to unmeasured cations (for example, ammonium and organic additives in the fracking fluids), and has been documented in other studies[33].

Marcellus and Utica fluid samples (paired to our metagenomic samples) were sent to the Pacific Northwest National Laboratory for NMR metabolite analysis. For the Utica input samples, technical duplicates for NMR analyses were highly similar, with mean concentrations reported (Supplementary Table 1). Samples were diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 (DSS) 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 analysed using Chenomx NMR Suite 8.1 with quantification based on spectral intensities relative to the internal standard. Candidate metabolites present in each of the complex mixtures 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[34], using a 1D nuclear Overhauser effect spectroscopy (NOESY) presaturation experiment with 65,536 complex points and at least 512 scans at 298 K. Additionally, 2D spectra (including ^{1}H-^{13}C heteronuclear single-quantum correlation spectroscopy (HSQC), ^{1}H-^{1}H total correlation spectroscopy (TOCSY)) were acquired on most of the fluid samples, aiding in the 1D ^{1}H assignments of acetate, ethanol, ethylene glycol, methanol and MMA.

Because of its significance in this work as an intermediate linking GB fermentation to methanogenesis, MMA was further confirmed in a series of 1D ^{1}H NOESY and 2D ^{1}H-^{13}C HSQC NMR spectra, where 'spiking' of several different samples was made using an MMA standard. Two additions of ~25 μM MMA were made to fluid samples, and only the assigned MMA peak (^{1}H chemical shift ~2.62 ppm and ^{13}C chemical shift ~27.7 ppm) increased in intensity. GB concentrations were too low for confirmation with 2D NMR experiments in the produced fluids. The GB in the Marcellus produced fluid sample series was resolvable and quantified by only the ~3.30 ppm ^{1}H resonance but not at ~3.92 ppm due to spectral overlap with ethanolamine. This was confirmed by spiking using a GB standard. In the Utica produced fluids, both resonances (~3.27 and ~3.92 ppm ^{1}H) were overlapped with other resonances and could not be resolved by GB spiking. GB was quantifiable in enrichment cultures with Utica fluids supplemented with GB (10 mM), and in controls not amended with GB, largely due to the dilution of ethanolamine and other confounding compounds. To compare temporal trends in the key metabolites between the Marcellus and Utica input and produced fluids in Fig. 2, each metabolite concentration was graphed in R over time for both wells with the same x axis (time, days).

Metagenomic sequencing and assembly. For genomic sample collection, 300–1,000 ml samples were concentrated onto 0.22-μm pore size polyethersulfone (PES) filters (Millipore, Fisher Scientific). Viruses were probably obtained on filters by flocculation with iron, which precipitated during the filtering process when the samples were first exposed to oxygen[34]. Total nucleic acids were extracted from the filter using the PowerSoil DNA Isolation kit (MoBio) for Marcellus fluids and a modified phenol chloroform nucleic extraction[35] for Utica fluids and enrichment cultures. Total cells with intact membranes were enumerated from unfiltered fluid samples for calibrated gate ranges on a Guava EasyCyte flow cytometer (EMD Millipore). Briefly, samples were fixed with 1% glutaraldehyde and stained with 0.1% SYBR Gold (Life Technologies), and quantified via flow-cytometry. For each time point, technical triplicates were measured and the data reported in Fig. 1 represent the mean ± s.d. (n = 3).

For the Marcellus input and produced fluids, Illumina HiSeq 2000 libraries were prepared using the Nugen Ovation Ultralow Library System following the manufacturer's instructions. Genomic DNA was sheared by sonication, and fragments were end-repaired. Sequencing adapters were ligated and library fragments were amplified with ~8–10 cycles of PCR before Pippin Prep size selection, library quantification and validation. Libraries were sequenced on the Illumina HiSeq platform and paired-end reads of 113 cycles were collected. Fastq files were generated using CASSAVA 1.8.2. Similar protocols were used for the Utica fluid GB enrichment culture, where sequencing was conducted on an Illumina HiSeq 2500 platform using a Kapa Hyper Prep library system with five cycles of PCR before solid-phase reversible immobilization (SPRI) size selection.

All metagenomics methods and scripts contributing to analyses in this manuscript are included in Supplementary Data File 4. Briefly, Illumina sequences from each of the five samples (input, T7, T13, T82 and T328) were first trimmed from both the 5′ and 3′ ends using Sickle (https://github.com/najoshi/sickle), then each sample was assembled individually using IDBA-UD (refs 36,37) with default parameters. Scaffold coverage was calculated by mapping reads back to the assemblies using Bowtie2 (ref. 37). Given the dominance and high strain variation in some samples, highly abundant genomes (>400×) often failed to assemble. Using an approach outlined in ref. 38, subassemblies were performed to reconstruct the dominant genomes in the day 13 and day 82 samples, using 10 and 8% of the reads, respectively. Results from the subassemblies are included (Supplementary Table 2).

Metagenomic annotation and genomic binning. All scaffolds ≥5 kb (≥1 kb for subassemblies, *Methanohalophilus*-1, and the GB enrichment culture) were included when binning genomes from the metagenomic assembly. Scaffolds were annotated as described previously[36,37] by predicting open reading frames using MetaProdigal[39]. Sequences were compared using USEARCH[40] to KEGG, UniRef90 and InterproScan[41] with single and reverse best hit (RBH) matches greater than 60 bits reported. The collection of annotations for a protein were ranked: reciprocal best BLAST hits (RBH) with a bit score >350 were given the highest (A) rank, followed by reciprocal best blast hit to Uniref with a bit score >350 (B rank), blast hits to KEGG

with a bit score >60 (C rank), and UniRef90 with a bit score greater than 60 (C rank). The next rank represents proteins that only had InterproScan matches (D rank). The lowest (E) rank comprises the hypothetical proteins, with only a prediction from Prodigal but a bit score of <60. Complete annotation files for all contigs >1,000 are available for download from https://chimera.asc.ohio-state.edu/daly_et_al_nature.html.

Within each sample we obtained the genome resolved 'bins' using a combination of phylogenetic signal, coverage and GC content[36,37]. For each bin, genome completion was estimated based on the presence of core gene sets (highly conserved genes that occur in single copy) for Bacteria (31 genes) and Archaea (104 genes) using Amphora2 (ref. 42; Supplementary Table 4). Overages (gene copies >1 per bin) indicating potential misbins, along with GC and phylogeny, were used to manually remove potential contamination from the bins.

To illustrate the microbial similarities shared between three Marcellus wells located in close proximity, we used 16S rRNA gene membership and abundance data from our earlier 454 amplicon study[5] to generate a non-metric multidimensional scaling ordination using Bray–Curtis distances. The ordination had a stress of 0.12, indicating that the matrix data were well represented by the ordination (Fig. 1a). Samples selected for metagenomics reflect the changing community over time and are denoted by a star in Fig. 1a. The relative abundance of each assembled genome in a sample was calculated as a proportion of the summed average coverage of the binned contigs in each sample. The relative abundance of the taxa in the input fluid that become dominant at later time points (Fig. 1c) was based on the normalized relative abundance of reconstructed 16S rRNA genes using EMIRGE (ref. 11), as genomic bins were not recovered for all these taxa. To verify the accuracy of our binned genomes, near-full-length ribosomal 16S rRNA gene sequences were reconstructed from unassembled Illumina reads from Marcellus fluids and an Utica fluid GB enrichment culture using EMIRGE (Supplementary Data File 1)[11]. To reconstruct 16S rRNA gene sequences we followed the protocol with trimmed paired-end reads where both reads were at least 20 nucleotides used as inputs and 50 iterations. EMIRGE sequences were chimaera checked before phylogenetic gene analyses.

For some genomes that lacked high strain resolution, such as *Methanohalophilus*-1, we confirmed the manual binning using an emergent self-organizing map (ESOM) using both the metagenomic data and isolate genomes from *Marinobacter*, *Methanohalophilus*, *Methanolobus*, *Halanaerobium* and *Halomonas* isolated species, as described previously[36,43]. Tetranucleotide frequencies were calculated for ≥5 kb fragments, with the number of tetranucleotides in each fragment normalized on the basis of the total number of observations in all fragments, with these values robust Z-transformed. The resulting matrix was used to train an ESOM for 30 epochs using scripts previously reported (https://github.com/tetramerFreqs/Binning). The ESOM was visualized using the Databionic ESOM Tools software.

Taxonomic placement of the genome bins relied on the phylogenetic analyses of 16S rRNA and/or ribosomal proteins. To determine if the same genome was present in different time points, we calculated ANI values (http://enve-omics.ce.gatech.edu/ani/) using a two-way ANI, with ≥99% ANI considered an initial cutoff for identical genomes through time. For genera with ≥99% ANI values (*Arcobacter*, *Halanaerobium* and *Idiomarina*), we then aligned contigs to examine synteny using the progressive Mauve aligner[44] in Geneious R8 (ref. 45). If a clustered regularly interspaced short palindromic repeat (CRISPR) array was present in each genome being compared over time, the contigs with CRISPR arrays were preferentially chosen for alignment, as CRISPR arrays are hyper-variable regions and are dynamic at short timescales due to new spacer incorporation. Several high-quality bins (with a representative of each taxa persisting in later time points) were selected for manual curation and genome finishing.

Viral genome binning, CRISPR identification and links to microbial hosts. Viral contigs were identified through annotations by including contigs ≥5 kb with viral structural genes (for example, capsid proteins, tail proteins and terminases), contigs containing a high number of proteins with no known homology, and using Metavir 2 (ref. 46) comparison to the viral RefSeq database. Circular contigs, indicating complete viral genomes, were determined using two methods: (1) analysis in the Metavir 2 (ref. 46) software by identifying identical k-mers at the two ends of the sequence and (2) manually examining SAM files generated by Bowtie2 (ref. 47) for paired reads present at the two ends of the sequence at the appropriate coverage level. Similar contigs shared across time points, or between Utica and Marcellus shales, were identified by comparing contigs using the criteria of >95% ANI over 80% of the contig length, analogous to the clustering in ref. 48. Contigs within clusters were then individually aligned and manually inspected to confirm identical contig sequences.

Crass was used to identify CRISPR repeat and spacer sequences[49], and was run both on individual sample reads and combined reads from all samples. To identify the microbial hosts of viruses and determine if viral predation was ongoing in the deep shale, we used BLASTn with an E-value cutoff of 1e–8 to identify contigs with repeat and spacers. First, we matched repeats to genomic bins to identify host CRISPR loci. Next, within each identified host CRISPR loci, we matched the spacers identified by Crass to viral contigs, identifying the viruses (or highly similar viruses) the host had encountered previously. All matches were manually confirmed as perfect matches by aligning sequences in Geneious R8, and we used the CRISPR

Recognition Tool plugin (CRT, version 1.2) to confirm CRISPR loci in genomic bins. Spacer links between host genomic bins and viral contigs were used to construct a network (Fig. 3) in Cytoscape (version 3.1.0). Each spacer match between a host CRISPR loci and a virus represents an edge in the network; nodes represent hosts (ovals) or viruses (diamonds and circles). Multiple edges indicate cases where a host CRISPR loci had multiple spacer links to the same viral contig, and these are represented by multiple (curved) edges in the network. For identical genomes present across samples (for example, *Halanaerobium*-1) where the viral–host links were only detected in genomes from some samples, the genome in other samples (oval nodes without edges) was also included. These links were probably not detected due to low estimated genome completion in some samples. Based on a recent paper[50], we classified CRISPR-Cas modules by manually examining the operon architectures of annotated contigs. Although we cannot determine whether a particular viral contig was in a virulent/lytic state at the time of sampling, in the input and day 7 samples, the contigs with the highest read coverage, identified as viral by the above criteria, had coverage several-fold higher than the most abundant bacterial or archaeal contig (input, 5.1-fold higher; T7, 4.0-fold higher), suggesting these viruses were virulent at the time of sampling.

Phylogenetic and metabolic analyses. The 16S rRNA genes recovered from the five nearest neighbours to each EMIRGE sequence from our Marcellus metagenomic samples were obtained from SILVA (release 123)[51], anchored with cultivated representatives (Supplementary Date File 1). 16S rRNA gene sequences were aligned in Geneious R8 using MUSCLE, a phylogenetic tree was constructed with RAxML 7.2.8 (GTR Gamma nucleotide model, 999 bootstrap replicates), and relative abundance data over time were graphed using iTOL. For the S3 protein tree, amino-acid sequences were pulled from the Marcellus and Utica GB enrichment genomic bins and were augmented with sequences mined from NCBI and JGI IMG databases (Supplementary Data File 2). Sequences were aligned using MUSCLE version 3.8.31 and run through ProtPipeliner, a python script developed in-house for generation of phylogenetic trees (https://github.com/lmsolden/protpipeliner). A maximum likelihood phylogeny for the alignment of S3 ribosomal proteins was conducted using RAxML version 8.3.1 under the LG+α+γ model of evolution with 100 bootstrap replicates and visualized in iTOL. The 16S rRNA genes recovered for the key terminal taxa in Fig. 1 were also blasted to NCBI to show the relationship between key terminal taxa in this study and environmental sequences. Sequences were trimmed to the V3–V4 region, and the top ten non-redundant hits from NCBI were included for analysis. 16S rRNA gene sequences were aligned in Geneious R8 using MUSCLE, and a phylogenetic tree was constructed with RAxML 7.2.8 (GTR Gamma nucleotide model, 999 bootstrap replicates) (Supplementary Data File 3).

Metabolic profiling was largely conducted by manual analyses. For the osmoprotectants inventory, the annotated gene lists were searched by name, KEGG number and E.C. number with positive records saved to files that were manually inspected to remove misidentified genes. The results were compared to the same functions in available genomes from the same genus in the IMG database on 15 August 2015 (http://img.jgi.doe.gov/). For key functional genes we used both a list and homology-based approach to help annotate genes. The latter is important, as many methylamine cycling genes were incorrectly annotated or not included on scaffolds >5,000 bp. Putative GrdEGI/PrdA were identified from the Utica GB enrichment by blasting known homologues capable of glycine/sarcosine/GB/proline reduction. The MttB and GrdEGI/PrdA trees were constructed similarly to above by aligning manually curated amino sequences in MUSCLE with RAxML version 8.3.1 using the LG model of evolution with 100 bootstrap replicates. For key genes in the metabolic processes outlined in the main text, we confirmed protein structure and functional prediction via modelling, catalytic or structural residues, or phylogenetic analyses.

GB enrichment culture from Utica produced fluids. NMR metabolite analyses were performed on Utica produced fluids that had been filtered and stored at −80 °C since the time of collection, after 483 days of anoxic (100% N$_2$ headspace) incubation without amendment, and before and after amendment to stimulate GB utilizing and methane-producing organisms. The GB enrichment consisted of 30% anoxic, incubated, produced Utica fluid (day 96) and 70% sterile modified DSMZ 479 media dispensed in Balch tubes sealed with butyl rubber stoppers and aluminium crimps under an atmosphere of N$_2$/CO$_2$ (80:20, vol/vol). Before mixing with produced fluids, the modified DSMZ medium (per litre) included 87 g sodium chloride, 1.5 g potassium chloride, 6.0 g magnesium chloride, 0.4 g calcium chloride, 1.0 g ammonium chloride, 2.0 g yeast extract, 2.0 g trypticase peptone, 0.2 g coenzyme M, 0.2 g sodium sulfide, 4.0 g sodium bicarbonate and brought to a pH of 7.1 using 1 mM NaOH. After autoclaving, this medium was supplemented with 9.7 mM GB and a trace element solution (DSMZ 141). As a no-donor GB control, 30:70 produced fluid:medium was established in parallel that lacked GB substrate amendment. GB enrichment and control cultures were monitored for methane production over 32 days using a Shimadzu gas chromatograph equipped with a thermal conductivity detector (TCD) using helium as a carrier gas at 100 °C. At 32 days after initial enrichment, DNA was extracted for 16S rRNA gene bacterial and archaeal clone libraries (data not included) and metagenomics. Analyses were conducted as described above, with the exception that 1% of the reads were used to

recover the dominant *Halanaerobium* genomic bin. All scaffolds in the assembly, regardless of length, were searched for homologues for MttB, GrdI, and methanogenesis pathways, CRISPR arrays were only analysed for the binned scaffolds (nucleotide ≥1 kb).

References

1. US Energy Information Administration. *Technically Recoverable Shale Oil and Shale Gas Resources: An Assessment of 137 Shale Formations in 41 Countries Outside the United States.* Report No. DOE/EIA-0383ER (US EIA, 2013); http://www.eia.gov/analysis/studies/worldshalegas/archive/2013/pdf/fullreport_2013.pdf

2. Park, S. Y. & Liang, Y. Biogenic methane production from coal: A review on recent research and development on microbially enhanced coalbed methane (MECBM). *Fuel* **166,** 258–267 (2016).

3. Ravot, G., Casalot, L., Ollivier, B., Loison, G. & Magot, M. *rdlA*, a new gene encoding a rhodanese-like protein in *Halanaerobium congolense* and other thiosulfate-reducing anaerobes. *Res. Microbiol.* **156,** 1031–1038 (2005).

4. Akob, D. M., Cozzarelli, I. M., Dunlap, D. S. & Rowan, E. L. Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Appl. Geochem.* **60,** 116–125 (2015).

5. 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).

6. 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).

7. 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).

8. Waldron, P. J., Petsch, S. T., Martini, A. M. & Nüsslein, K. Salinity constraints on subsurface archaeal diversity and methanogenesis in sedimentary rock rich in organic matter. *Appl. Environ. Microbiol.* **73,** 4171–4179 (2007).

9. Wuchter, C., Banning, E. & Mincer, T. J. Microbial diversity and methanogenic activity of Antrim Shale formation waters from recently fractured wells. *Front. Microbiol.* **4,** 367 (2013).

10. Murali Mohan, A., Bibby, K. J., Lipus, D., Hammack, R. W. & Gregory, K. B. The functional potential of microbial communities in hydraulic fracturing source water and produced water from natural gas extraction characterized by metagenomic sequencing. *PLoS ONE* **9,** e107682 (2014).

11. 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 Biol.* **12,** R44 (2011).

12. Konstantinidis, K. T. & Rosselló-Móra, R. Classifying the uncultivated microbial majority: A place for metagenomic data in the *Candidatus* proposal. *Syst. Appl. Microbiol.* **38,** 223–230 (2015).

13. Vikram, A., Lipus, D. & Bibby, K. Produced water exposure alters bacterial response to biocides. *Environ. Sci. Technol.* **48,** 13001–13009 (2014).

14. Oren, A. in *The Prokaryotes—Prokaryotic Communities and Ecophysiology* (ed. Rosenberg, E.) 421–440 (Springer, 2013).

15. Gaston, M. A., Jiang, R. & Krzycki, J. A. Functional context, biosynthesis, and genetic encoding of pyrrolysine. *Curr. Opin. Microbiol.* **14,** 342–349 (2011).

16. Andreesen, J. R. Glycine reductase mechanism. *Curr. Opin. Chem. Biol.* **8,** 454–461 (2004).

17. Bicknell, B. & Owens, J. D. Utilization of methyl amines as nitrogen sources by non-methylotrophs. *J. Gen. Microbiol.* **117,** 89–96 (1980).

18. Levin, I., Meiri, G., Peretz, M., Burstein, Y. & Frolow, F. The ternary complex of *Pseudomonas aeruginosa* alcohol dehydrogenase with NADH and ethylene glycol. *Protein Sci.* **13,** 1547–1556 (2004).

19. Kivisto, A. et al. Genome sequence of Halanaerobium saccharolyticum subsp. saccharolyticum strain DSM 6643T, a halophilic hydrogen-producing bacterium. Genome Announcements 1, e00187-00113 (2013).

20. Kang, X.-M., Wang, F.-F., Zhang, H., Zhang, Q. & Qiana, W. Genome-wide identification of genes necessary for biofilm formation by nosocomial pathogen Stenotrophomonas maltophilia reveals that orphan response regulator FsnR is a critical modulator. Appl. Environ. Microbiol. 81, 1200–1209 (2015).

21. Liang, R., Grizzle, R. S., Duncan, K. E., McInerney, M. J. & Suflita, J. M. Roles of thermophilic thiosulfate-reducing bacteria and methanogenic archaea in the biocorrosion of oil pipelines. Front. Microbiol. 5, 89 (2014).

22. Teske, A. et al. Diversity of thiosulfate-oxidizing bacteria from marine sediments and hydrothermal vents. Appl. Environ. Microbiol. 66, 3125–3133 (2000).

23. Fathepure, B. Z. Recent studies in microbial degradation of petroleum hydrocarbons in hypersaline environments. Front. Microbiol. 5, 173 (2014).

24. Montes, M. J., Bozal, N. & Mercadé, E. Marinobacter guineae sp. nov., a novel moderately halophilic bacterium from an Antarctic environment. Int. J. Syst. Evol. Microbiol. 58, 1346–1349 (2008).

25. Singer, E. et al. Genomic potential of Marinobacter aquaeolei, a biogeochemical 'opportunitroph'. Appl. Environ. Microbiol. 77, 2763–2771 (2011).

26. Nyyssonen, M. et al. Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. ISME J. 8, 126–138 (2013).

27. Labonté, J. M. et al. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. Front Microbiol. 6, 349 (2015).

28. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709–1712 (2007).

29. Makarova, K. S. et al. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 9, 467–477 (2011).

30. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. 7, 10613 (2016).

31. Emerson, J. B. et al. Virus–host and CRISPR dynamics in archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. Archaea 2013, 370871 (2013).

32. Goris, J. et al. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91 (2007).

33. Chapman, E. C. et al. Geochemical and strontium isotope characterization of produced waters from Marcellus shale natural gas extraction. Environ. Sci. Technol. 46, 3545–3553 (2012).

34. Weljie, A. M., Newton, J., Mercier, P., Carlson, E. & Slupsky, C. M. Targeted profiling: quantitative analysis of ^{1}H NMR metabolomics data. Anal. Chem. 78, 4430–4442 (2006).

35. Wrighton, K. C. et al. A novel ecological role of the Firmicutes identified in thermophilic microbial fuel cells. ISME J. 2, 1146–1156 (2008).

36. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211 (2015).

37. Wrighton, K. C. et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337, 1661–1665 (2012).

38. Hug, L. A. et al. Aquifer environment selects for microbial species cohorts in sediment and groundwater. ISME J. 9, 1846–1856 (2015).

39. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. Bioinformatics 28, 2223–2230 (2012).

40. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).

41. Quevillon, E. et al. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120 (2005).

42. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28, 1033–1034 (2012).

43. Dick, G. J. et al. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10, R85 (2009).

44. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5, e11147 (2010).

45. Kearse, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649 (2012).

46. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15, 76 (2014).

47. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

48. Brum, J. R. et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348, 1261498 (2015).

49. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. 41, e105 (2013).

50. Makarova, K. S. et al. An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. 13, 722–736 (2015).

51. Quast, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41, D590–D596 (2012).

Acknowledgements

Author contributions

P.J.M. and K.C.W. designed the study. R.A.D., S.A.W., D.W.H. and P.J.M. collected the samples and performed geochemical and/or metabolite measurements. R.A.W. and K.C.W. assembled the metagenome data. R.A.D., M.A.B., D.N.M. and K.C.W. binned the metagenomes. M.A.B., D.J.K., R.V.T., D.N.M. and R.A.D. conducted phylogenetic analyses, and M.A.B., R.A.D., D.J.K., D.N.M., J.D.M., P.J.M. and K.C.W. contributed to the metabolic analyses. R.A.D., M.A.B., M.J.W., P.J.M. and K.C.W. integrated the data and drafted the manuscript. Constructive edits of the manuscript were provided by D.J.K., J.D.M. and J.A.K. All authors reviewed the results and approved the manuscript.

Additional information

Supplementary information is available for this paper. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.C.W.

Competing interests

The authors declare no competing financial interests.