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

Rebecca A. Daly1, Mikayla A. Borton1, Michael J. Wilkins1,2, David W. Hoyt3, Duncan J. Kountz1, Richard A. Wolfe4, Susan A. Welch2, Daniel N. Marcus1, Ryan V. Trexler4, Jean D. MacRae5, Joseph A. Krzycki1, David R. Cole2, Paula J. Mouser4 and Kelly C. Wrighton1*

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 shale 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 formation1. 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 recovery2, bacterial hydrogen sulfide production (‘reservoir souring’) decreases profits and contributes to corrosion and the risk of environmental contamination2. 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 others4–9 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 fluids10, 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 microbiologically 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 wells5. 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 were enriched after hydraulic fracturing. Here, the reconstruction of 31 unique genomes coupled to metabolite data from the Marcellus and Utica shale 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.
reflected the microbial identities and abundances in the unassembled reads, by comparing genome bin relative abundance to reconstructed near-full-length 16S rRNA genes11 (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, Q2 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 sets12, 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 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 study5 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 shaded in grey. 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 were 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 genes11 (Supplementary Data File 1).
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, 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, 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, glycine betaine (GB) was present in the fluids, but mannitol, sorbitol, ectoine 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.

Microbiologically 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 *Halanaerobium* genomes, uses a GB reductase (GrdHI), producing TMA and acetate via a Stickland reaction16 (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, www.img.jgi.doe.gov, April 2016). In the Marcellus shales, a persisting *Halanaerobium* (*Halan–; 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 points17. Although *Methanohalophilus* 16S rRNA genes have been reported from Antrim18 and Burket/Geneseo4 fractured shales, our genomic and metabolite findings identify the endogenous and exogenous sources of methylotrophic 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 of GB and other metabolites.
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\(^\text{(19)}\) (alcohol dehydrogenase), and fermented by Halanaerobium (propanediol dehydratase, acetaldehyde dehydrogenase) later to yield ethanol, hydrogen and acetate. Candidatus Frackibacter also has the capacity to produce acetate via GB fermentation, homoacetogenesis (H\(_2/\text{CO}_2\)) 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\(^\text{(46-59)}\). Our shale-hosted Halanaerobium genomes also have the capacity to ferment amino acids (for example, alanine, discussed in the methylamine section), 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\(^\text{(39)}\), these shale genomes encode genes for flagellar motility and cellular aggregation (for example, polysaccharide production and diguanylate cyclase)\(^\text{(20)}\) (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\(^\text{(3)}\)), imply a role for shale Halanaerobium in steel corrosion and reservoir souring\(^\text{(21)}\). 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\(^\text{(21)}\). 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\(^\text{(2)}\).

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\(^\text{(6)}\). We failed to recover any genes for anaerobic hydrocarbon degradation, but the near-complete Marinobacter and Halomonadaceae and 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\(^\text{(28)}\).
including Barnett\(^6\) and Marcellus\(^2\) 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 haxose 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, tail 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 I 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 adaptation 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 62 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 *grdf* (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\(^{22}\) 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 Halomonadaceae with detrimental sulfate 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 the ~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 shale 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.
discrepancy for the input samples is probably due to unmeasured cations (for example, ammonium and organic additives in the fracturing fluids), and has been documented in other studies. The 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, as described previously, by predicting open reading frames using MetaProdigal. Sequences were compared using USEARCH® to KEGG, UniRef90 and InterProScan with single and reverse best (RH) matches greater than 60 hits reported. The collection of annotations for a protein were ranked: reciprocal best BLAST hits (RH) 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 (D rank). The next rank represents proteins that only had InterproScan matches (D rank). The lowest rank represents proteins that were not matched to any of the above databases. The resulting matrix was used to train an ESOM for 30 epochs using scripts previously reported (https://github.com/tetramerFreq/Binning). The ESOM was visualized using the Databionic ESOM Tools software.

Metagenomic sequencing and assembly. For genomic sample collection, 300–1,000 ml samples were concentrated onto 0.22-µm pore size PES filters (Millipore, Fisher Scientific). Viruses were probably obtained on filters by flocculation with ferric iron, which precipitated during the freezing process when the samples were first exposed to oxygen. Total nucleic acids were extracted from the filters using the PowerSoil DNA Isolation kit (MoBio) for Marcellus fluids and a modified phenol chloroform nucleic extraction25 for Utica fluids and enrichment cultures. Total cells with intact membranes were enumerated from unprocessed fluids, both resonances (~3.27 and ~3.92 ppm) were overlapped with other resonances and could not be resolved by GB spiking.

Q5

Q6

Q7

Q8

Metagenomic annotation and genomic binning. All scaffolds ≥2 kb (~1 bp for subassemblies, Methanohalophilus-1, and the GB enrichment culture) were included when binning genomes from the metagenomic assembly. Scaffolds were annotated as described previously, by predicting open reading frames using MetaProdigal. Sequences were compared using USEARCH® to KEGG, UniRef90 and InterProScan with single and reverse best (RH) matches greater than 60 hits reported. The collection of annotations for a protein were ranked: reciprocal best BLAST hits (RH) 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 (D rank). The next rank represents proteins that only had InterproScan matches (D rank). The lowest rank represents proteins that were not matched to any of the above databases. The resulting matrix was used to train an ESOM for 30 epochs using scripts previously reported (https://github.com/tetramerFreq/Binning). The ESOM was visualized using the Databionic ESOM Tools software.

The genome placement bins relied on the phylogenetic analyses of 16S rRNA genes and ribosomal proteins. To determine if the same genome was present in different time points, we calculated ANI values (http://envi-mics.cce.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 these reference alignment axes. An ANI ≥99% was used for the reference alignment. The interlaced short palindromic repeat (CRISP) array was present in each genome being compared over time, the contigs with CRISP arrays were preferentially chosen for alignment, as CRISP 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. CRISP identification and links to microbial hosts. Viral contigs were identified through annotations by including contigs ≥2 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 individually aligned and manually inspected to confirm identical contig sequences.

Cras was used to identify CRISP repeat and spacer sequences4, 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 CRISP loci. Next, within each identified host CRISP loci, we matched the spacers identified by Crass to viral contigs, identifying the viruses (or highly similar viruses) that had infected the host. Phylogenetic 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 mtfB, gdfI, and methanogenesis pathways. CRISP arrays were only analysed for the binned scaffolds (≥1 kb).

Accession codes. Sequencing data have been deposited in the NCBI sequence read archive under Bioproject PRJNA3083826. The near-complete representative genomes from Candidatus Frackbacter-2, Halanaerobium-1, Halomonadaceae-1, Idiomarina-1, Marinobacter-3 population, Methanohalophilus and Methanohalovibrio were assigned accession numbers SAMN04432553, SAMN04417677, SAMN04432552, SAMN04432539, SAMN04432574, and SAMN04432769 and SAMN04432770, respectively. The Methanohalophilus and Halanaerobium genomes recovered from GB enrichment have been assigned accession numbers SAMN05172267 and SAMN05172290. The 16S rDNA 454 pyrotags from our previous study can be accessed from the NCBI under Bioproject accession number PRJNA229085, with biosample numbers SAMN02414908 to SAMN0241927.

Additionally, genomic information (annotation, nucleotide and amino-acid files for each genome listed above), the FASTA files used in any phylogenetic analyses and viral contigs, and all EMIRGE 16S rDNA sequences are provided at https://chimera.asc.aco.state.edu/daily_et_al.nature.html.

Received 17 May 2016; accepted 15 July 2016; published xx xx 2016

References


7. Murali Mohan, A. Shale%20Oil%20Resource%20Assessment.pdf


25. The authors declare no competing financial interests.
**Journal:** Nature Microbiology  
**Article ID:** nmicrobiol.2016.146  
**Article title:** Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales

<table>
<thead>
<tr>
<th>Query Nos.</th>
<th>Queries</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Please check that the sentence beginning “Here” is OK as amended</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>I have now expanded ANI here – please confirm.</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>Please clarify what you mean by “the methylamine section” – are you referring to the Supplementary Information?</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>Is there something missing after Marcellus here?</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>Please expand NOESY</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>Please expand HSQC and TOCSY</td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>Please expand PES</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>Is there an expanded form of IDBA-UD we can give?</td>
<td></td>
</tr>
<tr>
<td>Q9</td>
<td>Please expand SSU</td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>Please expand ZT</td>
<td></td>
</tr>
<tr>
<td>Q11</td>
<td>Please check that the layout of the text in parentheses is OK as presented</td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>Please expand TCD</td>
<td></td>
</tr>
<tr>
<td>Q13</td>
<td>Ref 1 – please check the inserted web link</td>
<td></td>
</tr>
<tr>
<td>Q14</td>
<td>Ref 4 – please check inserted volume number and page range</td>
<td></td>
</tr>
<tr>
<td>Q15</td>
<td>Ref 9 – please check inserted volume number and article number</td>
<td></td>
</tr>
<tr>
<td>Q16</td>
<td>Ref 23 – please check inserted volume number and article number</td>
<td></td>
</tr>
<tr>
<td>Q17</td>
<td>Ref 27 – please check inserted article number</td>
<td></td>
</tr>
<tr>
<td>Q18</td>
<td>Ref 29 – please check amended journal name</td>
<td></td>
</tr>
<tr>
<td>Q19</td>
<td>Ref 30 – please check inserted article number</td>
<td></td>
</tr>
<tr>
<td>Q20</td>
<td>Ref 31 – please check inserted article number</td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Q21</td>
<td>Ref 36 – please check inserted volume number and page range</td>
<td></td>
</tr>
<tr>
<td>Q22</td>
<td>Ref 50 – please check amended journal name</td>
<td></td>
</tr>
<tr>
<td>Q23</td>
<td>Should M.D.J. be J.D.M. here (there is no author M.D.J.)?</td>
<td></td>
</tr>
<tr>
<td>Q24</td>
<td>Please clarify what PI indicates here</td>
<td></td>
</tr>
<tr>
<td>Q25</td>
<td>Please check that the acknowledgements section is OK as amended</td>
<td></td>
</tr>
<tr>
<td>Q26</td>
<td>In Figure 3 there are two circles with numbers in – should these be diamonds?</td>
<td></td>
</tr>
</tbody>
</table>