PNAS
www.pnas.org

Supplementary Information for

In vitro interactions scaled to *in situ* conditions: microorganisms predict field scale biogeochemistry in hydraulically fractured shale

Mikayla A. Borton[1], David W. Hoyt[2], Simon Roux[3], Rebecca A. Daly[1], Susan A. Welch[4], Carrie D. Nicora[2], Samuel Purvine[2], Elizabeth K. Eder[2], Andrea J. Hanson[5], Julie M. Sheets[4], David M. Morgan[4], Shikha Sharma[6], Timothy R. Carr[6], David R. Cole[4], Paula J. Mouser[5], Mary S. Lipton[2], Michael J. Wilkins[1,4], and Kelly C. Wrighton[1*]

Kelly C. Wrighton
The Ohio State University
Department of Microbiology
484 W. 12th Avenue
440 Biological Sciences Building
Columbus, OH 43210
kwrighton@gmail.com

This PDF file includes:

Supplementary text
Figs. S1 to S11
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Datasets S1 to S10

Supplementary Information Text

Osmoprotection Activity

Hydraulic fracturing creates a unique environment for life in which salinities increase from freshwater to brine through time. In Utica well 1 sampled, salinities reached up to 95 g/L chloride (Dataset S1). The pressure of salinity is reflected in the microcosm microbial community, as the genomes recovered are halotolerant and have the genomic potential to cope with elevated osmolarity (Figure S5). Further, both mechanisms for osmoprotection, including the salt-in strategy and production of compatible solutes, are detected in the metaproteome (1-2) (Figure S5).

Halanaerobium use the salt-in strategy, specifically utilizing multiple copies of sodium/proton antiporters (nhaC) that regulate intracellular sodium concentration while also balancing the number of protons in the cell (1). Contrary to prior reports (1), *Halanaerobium* in this microcosm also actively import and synthesize known osmoprotectants (Figure S5). We show that choline, proline, and glutamine are being actively imported or synthesized by *Halanaerobium* with no mechanism for degradation (Figure S5). While proline and glutamine could be assimilated by the cell, choline is imported and neither degradation mechanism (choline lyase or choline dehydrogenase) is present in the proteome or genome. Furthermore, no published *Halanaerobium* genomes, isolates or from metagenomics, have the genomic potential to degrade choline. *Halanaerobium* strains can also import maltose and trehalose via ABC transporters, but these compounds are actively degraded to D-glucose by maltose phosphorylase and alpha, alpha-trehalose phosphorylase, respectively. Similarly, *Halanaerobium* can transport and degrade GB, making it an unlikely osmoprotectant. It should be noted that taurine and mannitol are likely not being imported into the cell because these compounds are not detected in the microcosm or in Utica produced fluids. Likewise, the respective transporters are not substrate specific and are able to import GB and fructose, respectively (Figure S5).

Like *Halanaerobium*, *Ca.* Uticabacter, employs both the salt-in and compatible solute strategy simultaneously. *Ca.* Uticabacter utilizes sodium/ proton antiporter (nhaC) and uptakes GB, proline, and glutamine, but does not degrade these compounds, suggesting that it is using them for osmoprotection (Figure S5). It is also possible that proline and glutamine are being used in protein synthesis. *Methanohalophilus* actively imports and synthesizes GB for osmoprotection from glycine and sarcosine by glycine and sarcosine methyltransferases, respectively (Figure S5). *Geotoga* uptakes trehalose, maltose, glutamine and GB, with GB and the sugars being the likely compatible solutes. Maltose and trehalose are being interconverted via maltose alpha-D-glucosyltransferase by *Geotoga* but the proteins for degradation are not detected. Notably, the only two osmoprotectants detected in the produced fluids from the Utica well time series were GB and choline, suggesting that these two amines are key in microbial salinity tolerance (Dataset S1). Furthermore, based on GB trends, we infer both utilization (for osmoprotection, energy generation, and carbon and nitrogen assimilation) and production (for osmoprotection) of GB.

Viruses

Viruses accounted for 0.9% of the total microcosm metagenomic reads, denoting their prevalence in this *in vitro* ecosystem. Notably, viral peptides were detected in the

metaproteomics data. Our microcosom proteomic data also provided evidence for the activity of both viral lifestyles. Evidence for virion-producing active infections, as opposed to a lysogenic state, was provided by detection of multiple peptides for capsid production (e.g. terminase and head proteins). Also, we have evidence that some viral members are entering the lysogenic cycle and integrating themselves into host genome, as viral recombinase and resolvase proteins were also expressed.

Using nucleotide frequency (3), we demonstrated that viruses could be associated with every host. We predicted *Halanaerobium* was the most likely host for 8 viruses, *Methanohalophilus* for 2 viruses*, Ca.* Uticabacter for 4 viruses, and Geotoga for 2 viruses. Coordinated host and viral abundance patterns over time revealed no significant differences due to GB amendment, suggesting this treatment had little impact on viral predation. For three of the four microbial members, the microcosm viruses exhibited the same dynamics as their hosts over time (Figure S6). Alternatively, for *Methanohalophilus* and its most abundant associated virus, there was a clear decoupling between host and virus abundance patterns over time regardless of amendment.

To more directly link host and viral population genomes, we performed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) array analysis. Two of our microbial members had CRISPR arrays, with *Methanohalophilus* encoding 148 spacers in two CRISPR arrays and *Halanaerobium* encoding 206 spacers in four CRISPR arrays encoded a CRISPR-Cas system. None of the spacer sequences within the *Methanohalophilus* arrays matched viral genomes in our microcosms, suggesting that these spacers likely reflected historical viral encounters. We were able to link 8 *Halanaerobium* spacer sequences to 4 microcosm viral populations. Additionally, 14 of these Halanaerobium spacers also linked to 8 viruses recovered in the well used for this inoculum, as well as 2 viruses from another previously published well (4).

To predict if viral predation was ongoing in microcosms, we examined hosts for expression of CRISPR-Cas immunity genes. CRISPR-Cas proteins for the three functional stages of adaptive immunity were detected (adaptation, expression, interference) (Figure 2). Both *Methanohalophilus* and *Halanaerobium* expressed proteins for the adaptive stage of immunity (Cas1), suggesting active incorporation of spacers into the CRISPR loci. Concurrently, both also expressed proteins for the interference stage of viral immunity. *Halanaerobium* proteins were detected for the first part of the interference stage (Cas 5), which is implicated in producing cognate RNA that binds to invading DNA (5). Alternatively, *Methanohalophilus* proteins were detected for both parts of the interference stage (Cas 5 and Cas3), including cleavage of foreign viral DNA (5). Proteins for the expression stage (Cas6) were only detected from *Methanohalophilus* (Figure 2).

Stickland Reaction

The Stickland reaction is characterized by the oxidation of one amino acid coupled to the reduction of another (6,7). Since the discovery of this metabolism in 1934, several other non-amino acids have been characterized to take part in this reaction, including glycine betaine (GB), sarcosine, and ornithine (7-9). Organisms use this metabolism to generate energy in the form of ATP via substrate level phosphorylation (10). Several organisms have been described to take part in this reaction, most of them members of the class Clostridia (9-10). Here we describe an active Stickland reaction in *Halanaerobium* and *Candidatus* Uticabacter.

Both *Halanaerobium* and *Candidatus* Uticabacter in the microcosm experiment use reductase mechanisms related to the glycine reductase mechanism (9). This family of reductase systems can reduce glycine, sarcosine, proline, or GB. In each system, there are generally three proteins: protein A (encoded by *grdA*), protein B (encoded by different genes based on substrate specificity), and protein C (encoded by *grdCD*) (9). In the microcosm, *Halanaerobium* use the GB specific protein B (GrdHI*)*, while *Candidatus* Uticabacter use both the sarcosine and glycine specific protein B (GrdFG and GrdBE, respectively). Methods for determining reductase specificity were reported previously (4*)*. Briefly, alignments of the GrdE/I/G/PrdA homolog amino acid sequences from our metagenomic database and known GrdE/I/G/PrdA from *Eubacterium acidominophilum* and *Clostridium sticklandii* revealed that the *Halanaerobium* homolog lacked a conserved cysteine residue and formed a monophyletic clade with other known GB specific reductases (Figure S7). Furthermore, two *Candidatus Uticabacter* homologs clustered with known sarcosine and glycine reductases. Both organisms actively employ these reductase mechanisms, with all proteins detected in the proteome. For *Halanaerobium*, a bin likely composed of multiple strains, there are two full mechanisms turned on: grdA (scaffold_194_1, scaffold_93_1), grdH (scaffold_194_3, scaffold_93_3), grdI (scaffold_194_2, scaffold_93_2), grdC (scaffold_69_9, scaffold_93_9), and grdD (scaffold_69_8, scaffold_93_10). *Ca.* Uticabacter used a sarcosine reductase and a glycine reductase: grdA (scaffold_169_8, scaffold_23_32), grdB (glycine specific, scaffold_169_4), grdE (glycine specific, scaffold_169_3), grdG (sarcosine specific, scaffold_23_26), grdF (sarcosine specific, scaffold_23_27), grdC (scaffold_169_9, scaffold_23_33), and grdD (scaffold_169_10, scaffold_23_34). For *Candidatus* Uticabacter, all proteins were detected except for GrdA. Given that GrdA was detected in low amounts relative to the rest of the operon for the highly abundant *Halanaerobium*, we posit that *Candidatus* Uticabacter is likely using the sarcosine reductase and GrdA is just below detection. *Candidatus* Uticabacter using multiple reductase mechanisms has been found previously in other organisms including *C. sticklandii* and *C. difficile* (9). Moreover, this finding is consistent with the only other published genome from this genus (*Dethiosulfatibacter aminovorans* DSM 17477) has the genomic potential for three reductase mechanisms specific to glycine, sarcosine, and glycine betaine (Figure S7). One possible source of sarcosine is creatine through enzyme creatinease, which is expressed in *Ca.* Uticabacter.

Several reductants can be used to reduce GB, sarcosine and glycine. Here we show that *Halanaerobium* use lysine, serine, threonine, glycine, methionine, glutamate and alanine, while *Ca.* Uticabacter uses glutamate, leucine, phenylalanine, glycine and threonine. The oxidation of one amino acid in the Stickland reaction provides reducing power for the reduction of another amino acid. The key enzyme in the generation of this reducing power for each reductant follows: lysine (3,5-diaminohexanoate dehydrogenase, E.C. 1.4.1.11), serine (serine dehydratase, E.C. 4.3.1.17), threonine (threonine dehydratase, E.C. 4.3.1.19 and threonine dehydrogenase E.C. 1.1.1.103), glycine (glycine cleavage system), methionine (methionine gamma-lyase, E.C. 4.4.1.11), glutamate (glutamate dehydrogenase, E.C. 1.4.1.4), alanine (alanine dehydrogenase, E.C. 1.4.1.1), and leucine (leucine dehydrogenase, E.C. 1.4.1.9) (10).

These reductants could account for about 39% of GB reduced from T_0 to $T_{M,}$ with lysine (17%), serine (7.2%), threonine (3.8%), glycine (4.1%), and methionine (6.7%) of

GB reduction (Dataset S4). Although glutamate and alanine are likely reductants in the Stickland reaction with GB, as the respective dehydrogenases were detected in the *Halanaerobium* proteome, these were not apparent by metabolite analyses, suggesting that alanine and glutamate are being synthesized more quickly than *Halanaerobium* is oxidizing them (Dataset S1).

Lysine and GB are the most likely Stickland pair in the microcosm. Lysine is oxidized to acetate, butyrate and ammonia through crotonyl-CoA, with the key enzyme for the Stickland reaction being 3,5-diaminohexanoate dehydrogenase [Fonknechten, 2010 #113]. This enzyme is active concomitant with the GB reductase mechanism with the highest detection at T_M in the GB microcosm. Metabolites confirm the oxidation of lysine, as lysine is reduced by 93% overtime and accounts for 17.1% of GB reduction from T_0 to T_M (Figure 3). Moreover, butyrate is produced (8.13±0.5 μmoles) in a nearly 1 to 1 ratio with lysine loss (7.8±0.5 μmoles) from T_0 to T_F, congruent with lysine oxidation.

Glycine Cleavage System

As discussed previously, glycine is used as a Stickland oxidant (*Ca.* Uticabacter), a Stickland reductant (*Ca.s* Uticabacter and *Halanaerobium*), and also in osmoprotectant synthesis (*Methanohalophilus*) (Figure 4). This multi-enzyme complex oxidizes glycine to CO_2 and methylene-THF (11). Although the reaction can be ran in reverse, we hypothesize that *Halanaerobium* and *Ca.* Uticabacter are oxidizing the metabolite, freeing electrons to complete the Stickland reaction. Metabolites confirm this finding in the GB amended microcosm as 265.7±6.3 μmoles of glycine is depleted to 21.1±1.0 μmoles from T_0 to T_F. Moreover, we speculate that *Geotoga*, runs the glycine cleavage system in reverse, producing glycine because metabolites show glycine production from T_M to T_F in the no GB microcosm, when *Geotoga* activity is highest (Figure 1).

Ethanolamine Utilization

Halanaerobium employs a mechanism for ethanolamine utilization (Figure 4). Congruently, ethanolamine was detected in every time point of Utica produced fluids sampled here (Dataset S1). In the microcosm, *Halanaerobium* converts ethanolamine, present in the produced fluid inoculum, into acetaldehyde and ammonium by using the ethanolamine ammonia lyase (EutBC, 4.3.1.7) (Figure 4, Figure S9). Acetaldehyde is then converted into acetyl-aldehyde by the aldehyde oxidoreductase (EutE) and subsequently to acetate through acetylphosphate. Alternatively, acetaldehyde can be converted to ethanol by an alcohol dehydrogenase (EutG), which is often thought to be used as a detox mechanism (12) (Figure S9). Microcosm metabolites confirm this metabolism, as ethanolamine is reduced from a concentration of 165.3 ± 7.4 μmoles and 119.0 ± 27.4 μmoles to below detect in GB and no GB microcosms, respectively (Figure 4). In the both the GB and no GB enrichment, EutE is detected at higher levels than EutG, suggesting that *Halanaerobium* is using ethanolamine for energy, rather as a detoxification mechanism.

Halanaerobium-encoded detected proteins for ethanolamine utilization include: ethanolamine ammonia lyase large subunit (EutB, scaffold_31_26), ethanolamine ammonia lyase small subunit (EutC, scaffold_31_25), acetylaldehyde dehydrogenase (EutE, scaffold_31_22), alcohol dehydrogenase (EutG, scaffold_31_10), microcompartments/ carboxysome structural proteins (scaffold_31_14, scaffold_31_21,

scaffold_31_24), ethanolamine transporter (EutH, scaffold_31_13). All proteins detected in *Halanaerobium* proteome for the Eut operon are shown in Figure S9. Ethanolamine ammonia lyase is a vitamin B12 requiring enzyme, thus *Halanaerobium* imports this cofactor via transporters and does not make it *de novo*. We note that ethanolamine transporter protein EutH is detected in low levels and that ethanolamine is likely diffusing across the membrane concurrent with transport (12) (Figure S9).

Methane and Acetate Mass Balance Calculations

Given the importance of GB to hydraulically fractured shale organisms, both as a substrate and an osmoprotectant, and the presence of GB in the Utica well sampled, we amended produced fluids with GB and tracked microbial activity and metabolites through time (Figure 3). In the microcosm, *Halanaerobium* utilized GB reductase to reduce GB to TMA (*grdHI*), which was most active at T_M in the GB amended microcosm (Supplementary Information). Analysis of metabolites by NMR support the proteomics data, showing that in the GB amended microcosm 42.1±2.4 μmoles of GB was reduced to 37.9±0.6 μmoles of TMA from T_0 to T_M, a 90% reduction (Figure 3). Similarly, in the no GB microcosm, 2.6±0.1 μmoles of GB was 81.2% reduced to TMA from T_0 to T_M (Figure 3).

The TMA produced by *Halanaerobium* is utilized by *Methanohalophilus*, a methylotrophic methanogen (Figure 3). The most methane is produced from T_M to T_F in the GB amended microcosm (Figure 3). From T_M to T_F, 95% and 60% of TMA is converted to methane in the GB and no GB microcosm, respectively (Figure 3). Congruently, the most *Methanohalophilus* proteins are detected in T_F timepoints, with the GB amended microcosm having statistically more than the no GB microcosm. Furthermore, the trimethylamine specific pyrrolysine-containing methyltransferase (MttB) and the corresponding corrinoid protein (MttC) were highly detected in T_F in the GB amended microcosm, statistically more than in any other sample. Methyltransferase proteins specific to dimethylamine, monomethylamine, and methanol and all proteins necessary for methanogenesis were also detected (Dataset S1). Dimethylamine and monomethylamine concentrations followed the same pattern as trimethylamine, increasing from T_0 to T_M and decreasing in from T_M to T_F (Dataset S2). If we assume all methane production was fueled indirectly by GB, 72% of GB accounts for all methane produced in the GB amended microcosm from T_0 to T_F (Figure 3). There was no potential for GB or choline demethylation in our microcosm experiments, as no non-pyrrolysine trimethylamine methyltranserfases were detected (13-14).

Acetate, also produced one to one with TMA in the reduction of GB, had a net increase of 63.8±1.5 μmoles from T_0 to T_M in the GB amended microcosm (Figure 4). The excess acetate (25.9±1.5 μmoles) produced in the GB amended microcosm can be accounted for by residual carbon fermentation, as the no GB microcosm produced 24.0±1.7 μmoles acetate, of which only 2.5±0.1 μmoles came from GB fermentation. Given that acetate is produced in a one to one stoichiometric balance with TMA from GB reduction (9), we know that excess acetate (not accounted for by GB reduction, 25.9±1.5 μmoles) was produced in the GB amended microcosm. Notably this accounts for ~97% of acetate in non-amended microcosm (24.0±1.7 μmoles), where no GB was added. With GB accounting for 46% of acetate production, the excess can be

accounted for through sugar fermentation, with glucose (2.3%), trehalose (21.1%), ethylene glycol (11.1%), ethanolamine (1.4%), pyruvate (0.3%), maltose (2.3%), and fructose (4.0%) accounting a substantial portion of acetate production in the amended GB microcosm. See Dataset S1 for detailed acetate mass balance calculations.

Back to the field: Validation of microcosm generated hypotheses across wells

We compared our metabolic findings from microcosm experiments to previously published hydraulically fractured shale datasets and 33 metagenomes paired to metabolites published here. Prior to the Daly, *et al.* study, HF microbiology studies were limited to 16S rRNA analyses, did not have time series data including injected fluids, or did not include metabolites (15-17). Given that Daly *et al.* was a single well, it was necessary to apply our microcosm findings to other wells in different shale formations. Here we add 33 metagenomes and paired metabolites to build a HF database of 38 metagenomes. The 33 additional metagenomes came from injected fluids and produced fluids from four wells in the Marcellus and Utica shales. Two Utica wells were located in Ohio, two Marcellus wells in West Virginia, and one Marcellus well in Pennsylvania (Figure S10). Chloride concentrations increased over time in all wells (Figure S10). Metabolites and metagenome information can be found in Datasets S1 and S2, respectively.

In light of the importance of the Stickland reaction to hydraulically fractured shale organisms, we mined the published isolate genomes and metagenomes from produced fluids for the necessary genes (4,18-19). We found that 24% of genomes in our shale database had the potential to use GB, 5 of them *Halanaerobium*. Moreover, we found that the most abundant *Halanaerobium* strain at late time points in the well sampled here has a GrdI (*Halanaerobium* 6-U2, genome previously published in Booker, et al. (18). As previously reported, *Frackibacter*, a new genus within the Halobacteroidaceae discovered in shale, has the potential to reduce GB (4), and 2 of 3 publicly available *Frackibacter* genomes have the genomic potential to use GB.

SUPPLEMENTARY FIGURES

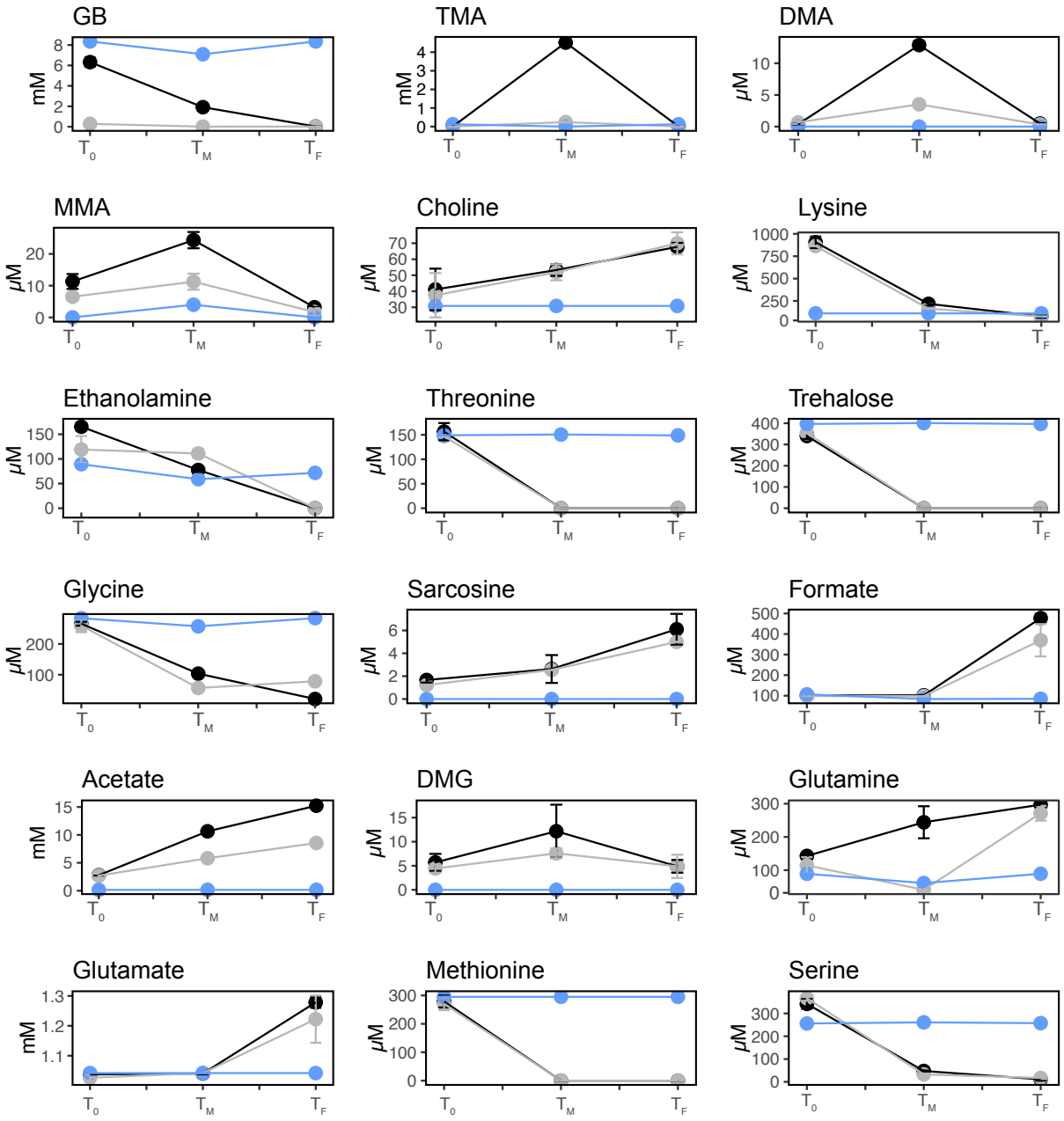


Fig. S1. Graphs show all metabolites detected in microcosms by NMR, with all treatments shown (with GB= Black, no GB= grey, Media Control= blue). Points indicate triplicate average and error bars show one standard deviation from the mean.

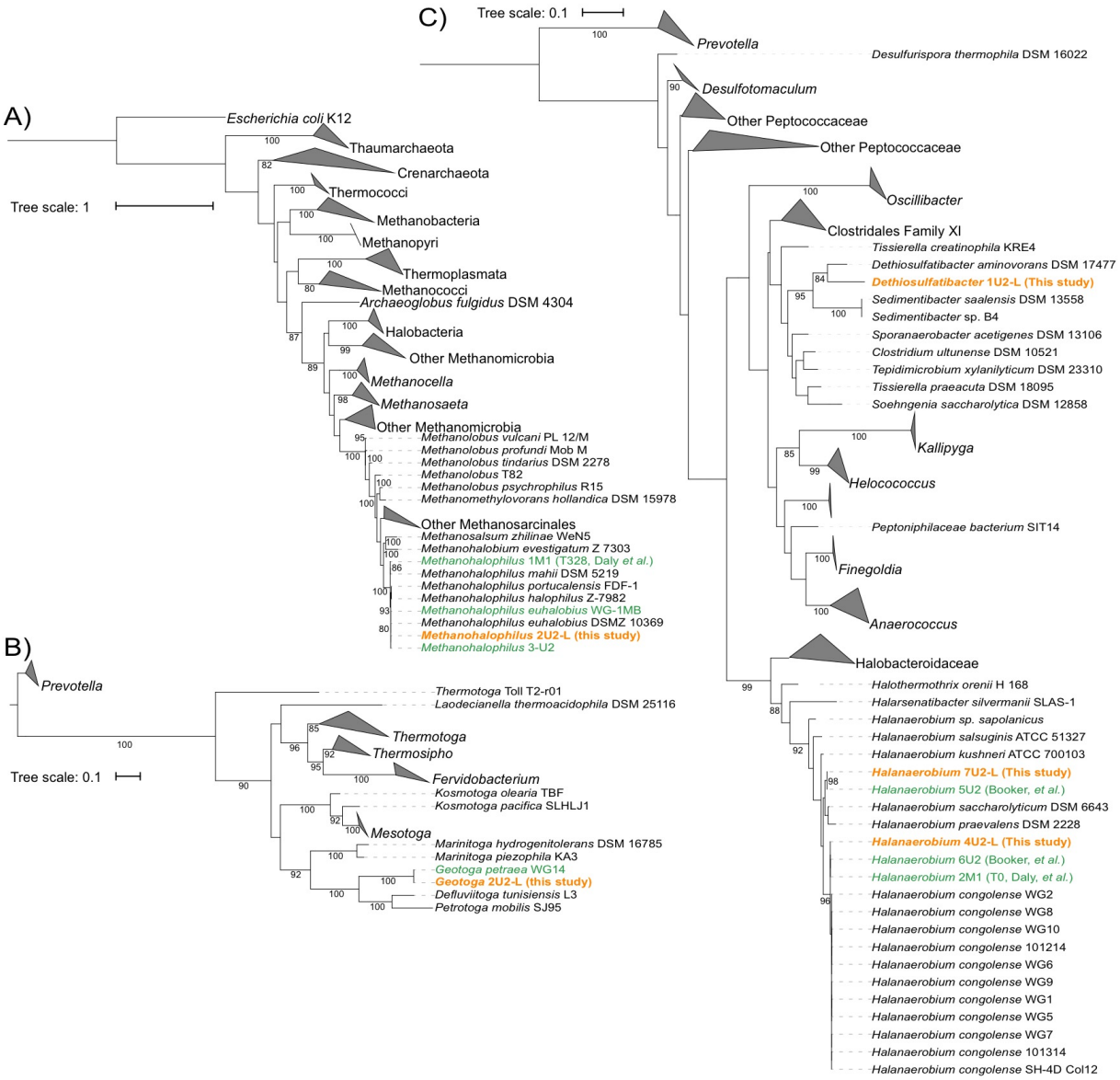


Fig. S2. Maximum likelihood S3 ribosomal protein trees of archaea (A) and Thermotogaceae (B), and Firmicutes (C), showing the taxonomic assignment of genomes from the microcosm experiment. S3 amino acid sequences from bins in this study are shown in orange, while sequences from (4) are shown in green.

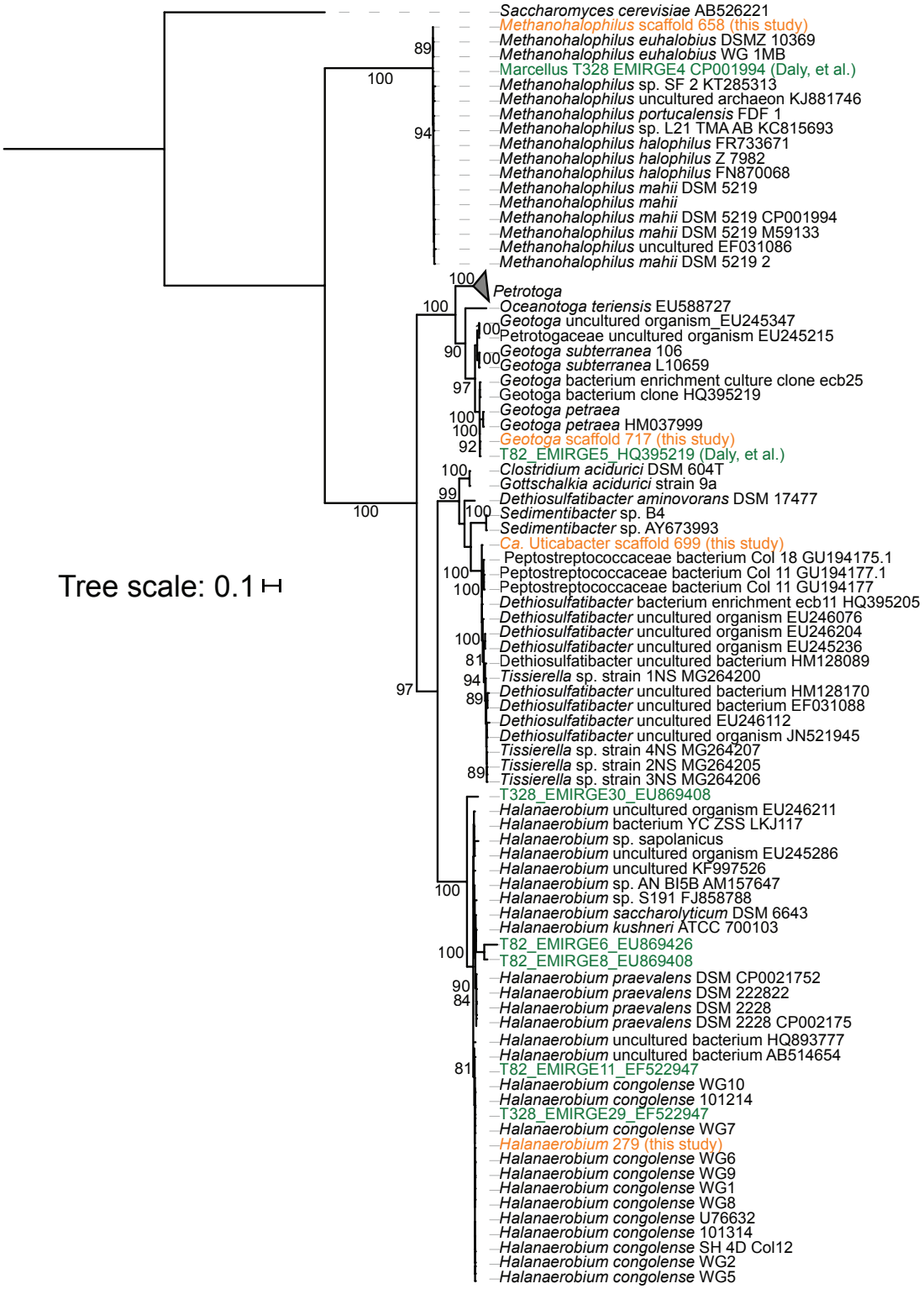


Fig. S3. Maximum likelihood 16S rRNA tree, showing the taxonomic assignment of genomes from the microcosm experiment. 16S rRNA from bins in this study are shown in orange, while sequences from (4) are shown in green.

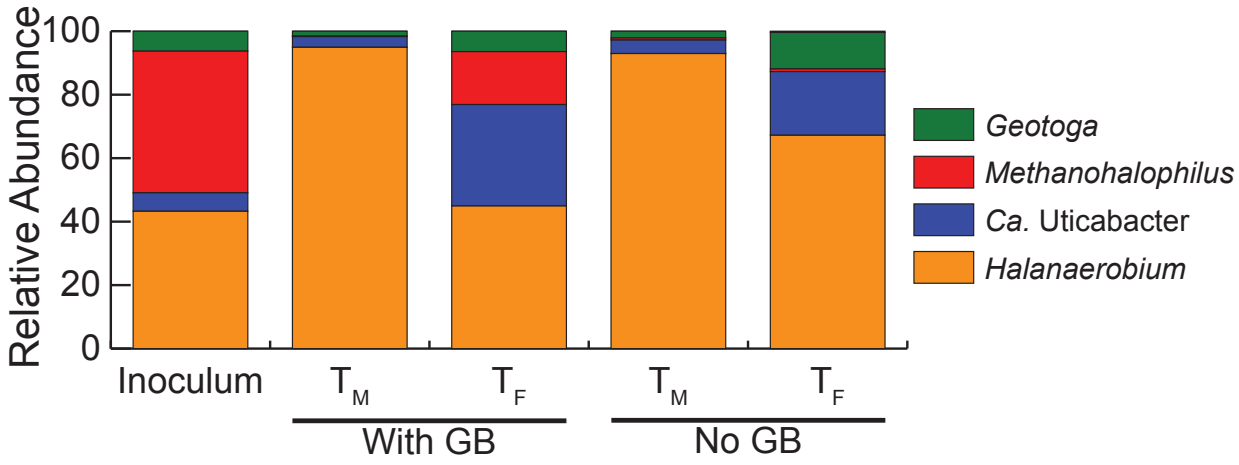


Fig. S4. Relative abundance by EMIRGE of all time points in the GB and non-GB microcosm experiments. Stacked bars are colored by organism within each metagenome. Organisms with >0.05% relative abundance are shown.

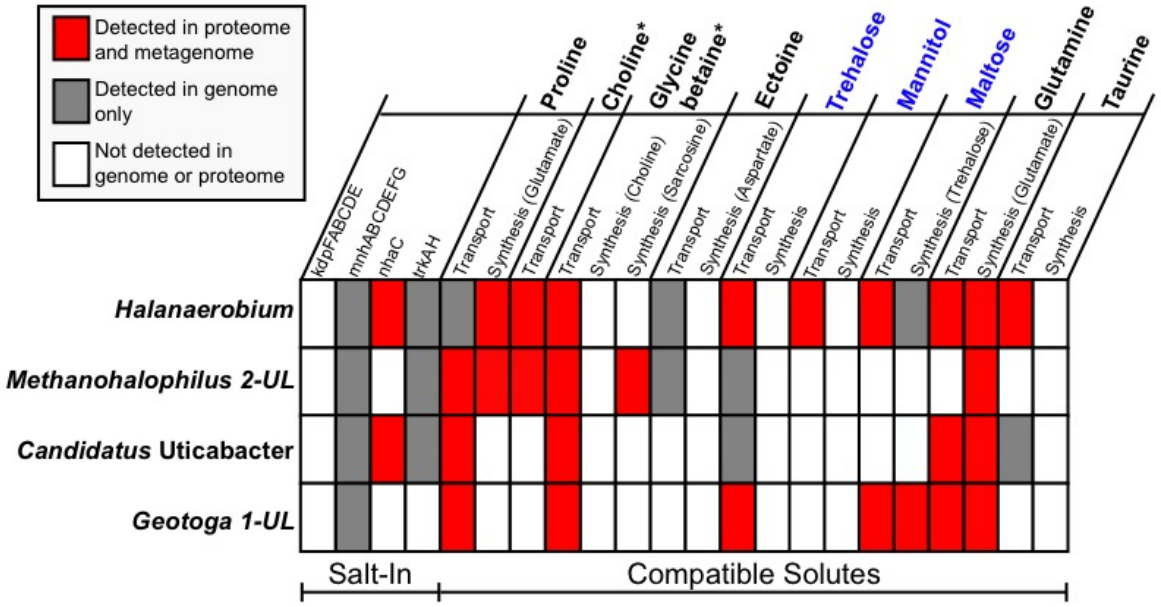


Fig. S5. Heatmap denotes active and potential osmoprotection strategies utilized by microcosm microbial community. Both salt-in and compatible solute strategies are considered. Compatible solute compounds found in the produced fluid Utica well time series are denoted with an asterisk (*), while sugar compatible solutes are shown in blue text. For multisubunit enzymes, >75% percent of proteins were required for detected in the proteome status.

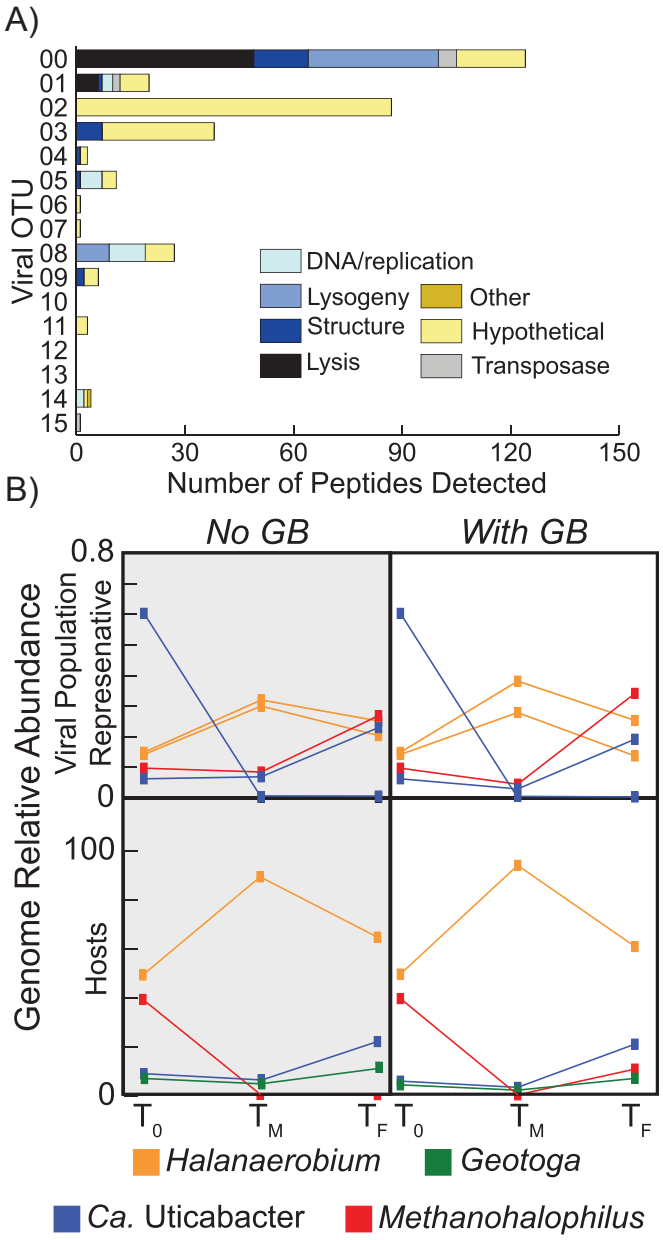


Fig. S6. A) Stacked bar chart denotes detected unique viral peptides per representative genome broken into 7 different categories (see methods). B) Genome relative abundance of microbial hosts (bottom) and viral population representatives (top) are shown across time and within treatments (No GB and With GB shown on the left and right, respectively). Only viral populations with >0.1% relative abundance in at least one timepoint in GB microcosm are shown. Viral OTUs represented in B include 00 (*Methanohalophilus*, red), 02 (*Ca.* Uticabacter, blue, decreasing from T_0 to T_F) ,10 (*Halanaerobium*, orange, least abundant), 13 (*Ca*. Uticabacter, blue, increasing from T_0 to T_F), and 15 (*Halanaerobium*, orange, most abundant).

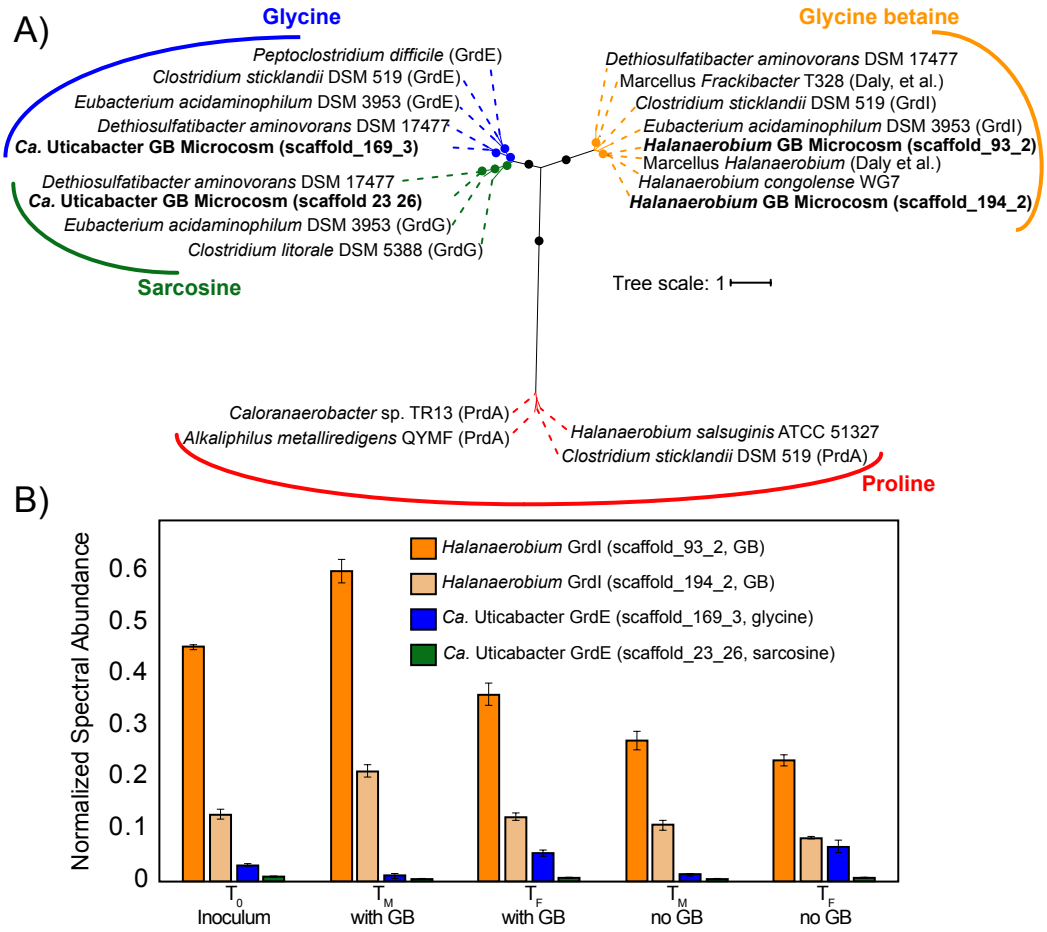


Fig. S7. Reductase systems for glycine, GB, and sarcosine are active in microcosm experiments. A) Phylogenetic analyses of GrdE (blue, glycine), GrdG (green, sarcosine), GrdI (orange, GB), and PrdA (red, proline) proteins from microcosm experiments showed that proteins clustered by substrate specificity. *Halanaerobium* had two active copies of GB reductase and *Candidatus* Uticabacter had an active glycine and sarcosine reductase, as these formed monophyletic clades with known GB, with known reducers of the respective methylamine substrates, *Eubacterium acidominophilum* and *Clostridium sticklandii*. Sequences from this study are in bold and include the genome and scaffold number followed by the relevant gene number(s). Bootstraps >90 are shown with closed circles at nodes. B) Activity shown of the reductases in A) is shown by time point (x-axis) with color denoting reductase mechanism. Bars represent average activity of biological triplicates with standard deviation shown (error bars).

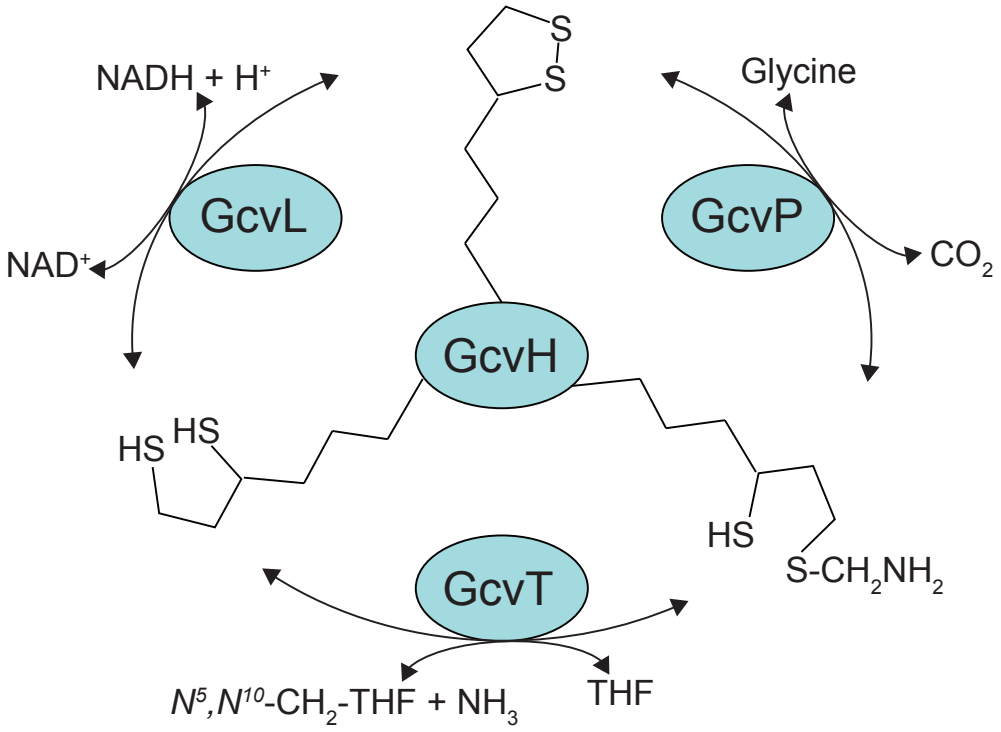


Fig. S8. Glycine cleavage system (Gvc) is shown, with four key proteins denoted in blue ovals.

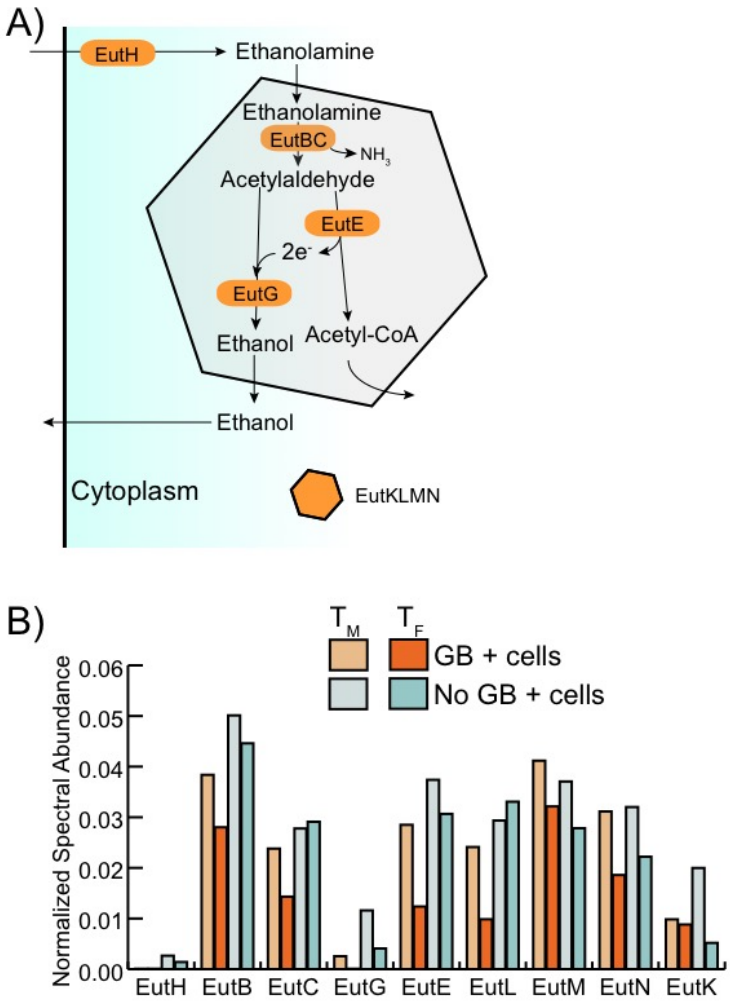


Fig. S9. Ethanolamine utilization in *Halanaerobium*. A) Pathway of ethanolamine utilization by *Halanaerobium.* All proteins shown were detected in the metaproteomics. B) Relative quantification of ethanolamine utilization proteins. Each bar represents the average NSAF value for each protein (in triplicate) within each time point by treatment.

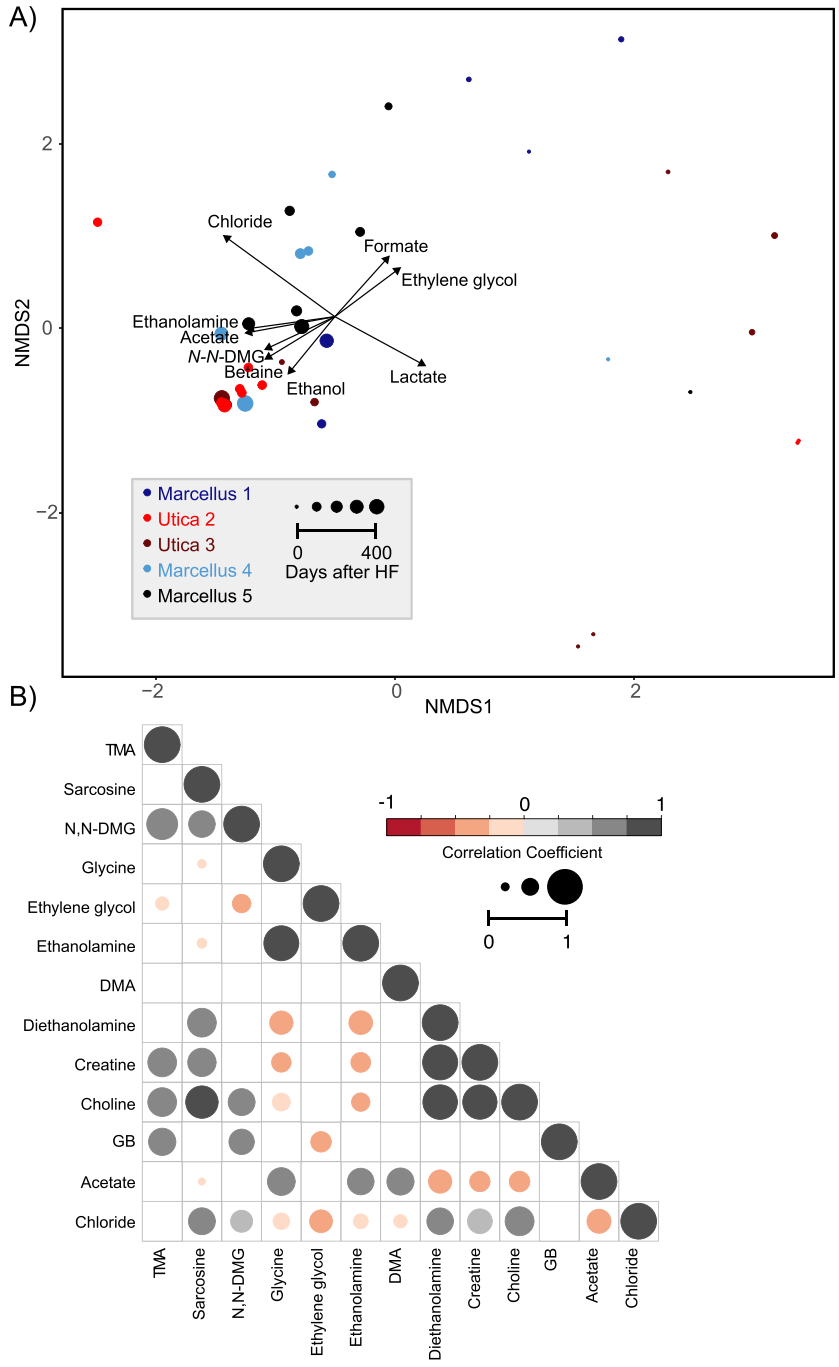
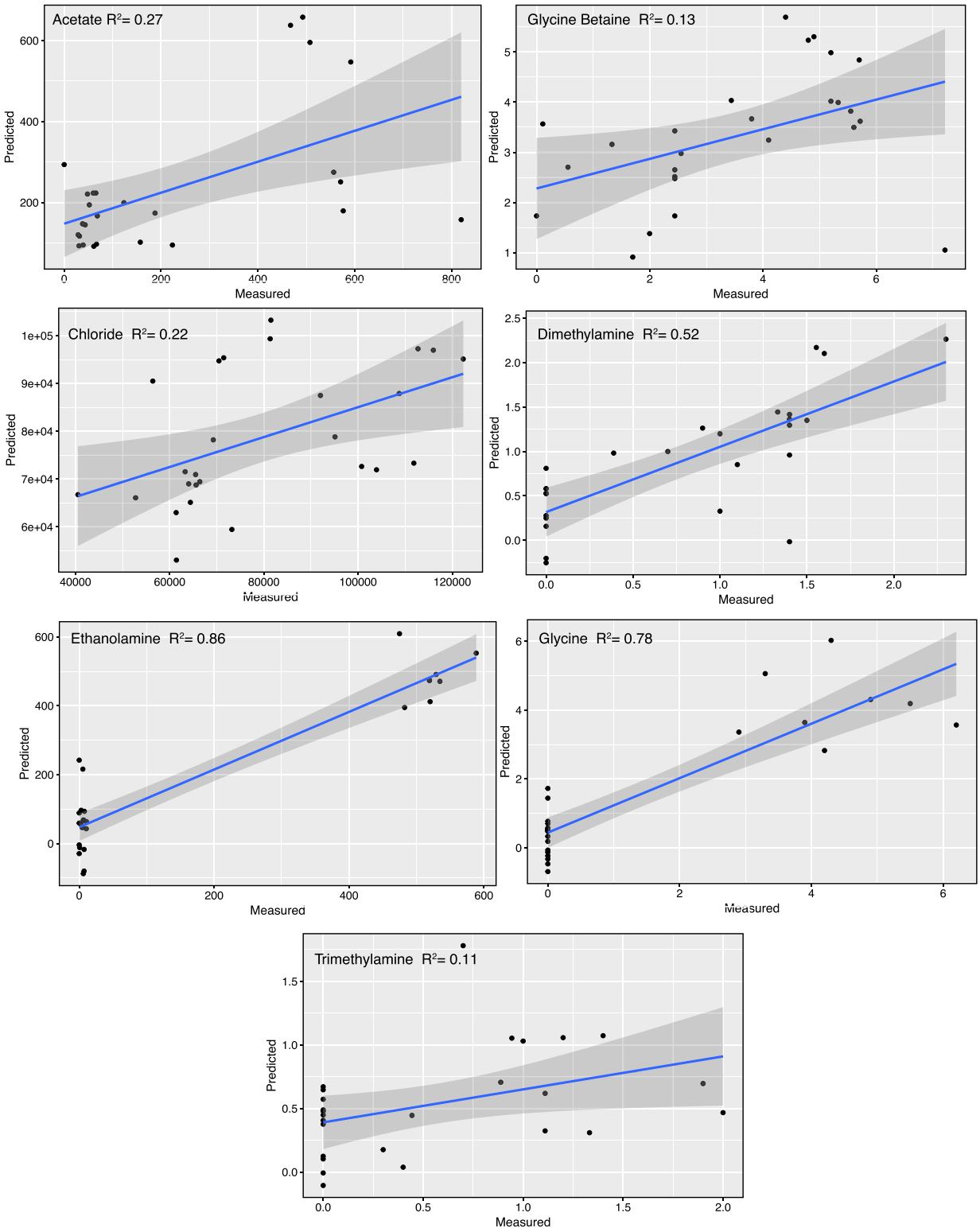


Fig. S10. A) Nonmetric multidimensional scaling of microbial community abundance overlaid with geochemistry. All vectors show significant associations between microbial communities and paired sample chemistry (envfit, p-value<0.05). Samples are colored by well and bubble size denotes time after HF. B) Bubble plot shows significant correlations between metabolites analyzed by NMR in 26 produced fluid samples collected from five HF wells, inputs were excluded from the analysis for clarity. Bubble color and size denotes correlation coefficients using colored scale bar below.



476
477 ***Fig. S11.*** Produced fluid microbial communities predict acetate, chloride, dimethylglycine,
478 glycine, trimethylamine, and glycine betaine using sPLS. Correlations of measured versus
479 predicted for each metabolite is shown, all p-values < 0.05.
480

Supplementary Dataset Legends:

Dataset S1 (Excel): Table of detected metabolites (μM) and chloride (mg/L) from field time series (*n=41*) collection and microcosm experiments (*n=21*). Concentrations of zero denote that the metabolite was below detection. Mass balance calculations are also included here.

Dataset S2 (Excel): Table optical density and gas chromatography measurements through time in the microcosm experiment.

Dataset S3 (Excel): Metagenomic and genomic data tables contain metagenome assembly statistics from field and microcosms, microbial and viral genome bin statistics, and scaffold and gene information for key metabolisms discussed.

Dataset S4 (Excel): File contains raw peptide data from metaproteomics, including the unique peptides per amino acid sequence (Protein Report tab), the peptides by amino acid sequence detected per sample (Proteins per sample tab), and the NSAF values (NSAF tab).

Dataset S5 (Excel): Strain resolved microbial abundances (by ribosomal S3 protein) across input and produced fluid samples.

Dataset S6 (.fasta): Nucleotide files of genome bins.

Dataset S7 (.fasta): Amino acid files of genome bins.

Dataset S8 (pdf): Maximum likelihood S3 ribosomal protein tree of unique S3 proteins from all HF metagenomes.

Dataset S9 (excel): Value Importance in Projection for each predicted metabolite in Figure 5.

Dataset S10 (Rscript): Rscripts used for and sPLS analyses.

References:

1. Oren A (1999) Bioenergetic aspects of halophilism. *Microbiology and molecular biology reviews* 63(2):334-348.
2. Sleator RD & Hill C (2002) Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS microbiology reviews* 26(1):49-71.
3. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, & Sun F (2016) Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research* 45(1):39-53.
4. Daly RA*, et al.* (2016) Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nature microbiology* 1:16146.

5. Makarova KS*, et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology* 9(6):467-477.
6. Stickland LH (1934) Studies in the metabolism of the strict anaerobes (genus Clostridium): The chemical reactions by which Cl. sporogenes obtains its energy. *Biochemical Journal* 28(5):1746.
7. Nisman B (1954) The stickland reaction. *Bacteriological reviews* 18(1):16.
8. Naumann E, Hippe H, & Gottschalk G (1983) Betaine: new oxidant in the Stickland reaction and methanogenesis from betaine and L-alanine by a Clostridium sporogenes-Methanosarcina barkeri coculture. *Applied and environmental microbiology* 45(2):474-483.
9. Andreesen JR (2004) Glycine reductase mechanism. *Current opinion in chemical biology* 8(5):454-461.
10. Fonknechten N*, et al.* (2010) Clostridium sticklandii, a specialist in amino acid degradation: revisiting its metabolism through its genome sequence. *Bmc Genomics* 11(1):555.
11. Andreesen JR (1994) Glycine metabolism in anaerobes. *Antonie Van Leeuwenhoek* 66(1-3):223-237.
12. Garsin DA (2010) Ethanolamine utilization in bacterial pathogens: roles and regulation. *Nature Reviews Microbiology* 8(4):290-295.
13. Ticak T, Kountz DJ, Girosky KE, Krzycki JA, & Ferguson DJ (2014) A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. *Proceedings of the National Academy of Sciences* 111(43):E4668-E4676.
14. Watkins AJ, Roussel EG, Webster G, Parkes RJ, & Sass H (2012) Choline and N, N-dimethylethanolamine as direct substrates for methanogens. *Applied and environmental microbiology* 78(23):8298-8303.
15. Mouser PJ, Borton M, Darrah TH, Hartsock A, & Wrighton KC (2016) Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS microbiology ecology* 92(11).
16. Cluff MA, Hartsock A, MacRae JD, Carter K, & Mouser PJ (2014) Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus Shale gas wells. *Environmental science & technology* 48(11):6508-6517.
17. Akob DM, Cozzarelli IM, Dunlap DS, Rowan EL, & Lorah MM (2015) Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Applied Geochemistry* 60:116-125.
18. Booker AE*, et al.* (2017) Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* 2(4):e00257-00217.
19. Lipus D, Vikram A, Ross DE, & Bibby K (2016) Draft genome sequence of Methanohalophilus mahii strain DAL1 reconstructed from a hydraulic fracturing-produced water metagenome. *Genome announcements* 4(5):e00899-00816.