Genomic Evidence for a Chemical Link Between Redox Conditions and Microbial Community Composition

Short title: Geobiochemistry of microbial proteomes and communities

Jeffrey M. Dick,[1*] Jingqiang Tan[1*]

[1]Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring, Ministry of Education, School of Geosciences and Info-Physics, Central South University, Changsha 410083, China

*Corresponding authors. Email: jeff@chnosz.net or tanjingqiang@csu.edu.cn

Abstract: Many microbial ecology studies use multivariate analyses to relate microbial abundances to independently measured physicochemical variables. However, genes and proteins are themselves chemical entities; in combination with genome databases, differences in microbial abundances therefore quantitatively encode for chemical variability. We combined community profiles from published 16S rRNA gene sequencing datasets with predicted microbial proteomes from the NCBI Reference Sequence (RefSeq) database to generate a two-dimensional chemical representation of microbial communities. This analysis demonstrates that environmental redox gradients within and between hydrothermal systems and stratified lakes and marine environments are reflected in predictable changes in the carbon oxidation state of inferred community proteomes. These findings have important implications for understanding

the microbial communities in produced well fluids and streams affected by hydraulically fractured wells. Although redox measurements for these environments generally are not available, this analysis suggests that redox chemistry is likely to be a significant driver of the microbial ecology of these systems.

Teaser

A chemical representation of predicted microbial proteomes reveals new links between community structure and environmental chemistry.

Introduction

A basic question for microbial ecology is how environmental conditions shape the taxonomic composition of microbial communities. In a common workflow, next-generation sequencing is followed by multivariate statistical analysis to cluster similar sequences into operational taxonomic units (OTU) (*1*). Then, patterns in OTU abundances are revealed using ordination methods and combined with physicochemical measurements to relate community changes to environmental parameters (*2*). Despite the power of ordination methods to uncover environmental signals in community sequence data, a well-known issue is the problem of unmeasured environmental variables, also known as latent factors. Spatial distance and even study ID (which may be related to geographic location) have been proposed as a proxy for unmeasured variables in some regional and global microbial biogeography studies (*3, 4*). However, a distance-based metric may be an inadequate proxy for physicochemical drivers. For instance, a principal components analysis (PCA) biplot for community structure at the Hawaii Ocean Time-series (HOT) ALOHA station showed clear separation of near-surface and deep-water communities; these clusters were distinguished by differences of temperature and bulk geochemical measurements but surprisingly not depth, which therefore appears to be a poor proxy for an obvious yet unmea-

sured driver of community structure, i.e. photon availability (*5*). The primary physicochemical driver of community structure on a global scale is thought to be salinity (*4, 6*), but some of the same workers have noted that oxygenation conditions may also be important (*6*). In particular, oxygen concentration or other measures of oxidation-reduction (redox) conditions are likely to have large effects of community structure in settings such as hydrothermal systems and oxygen minimum zones.

To better understand the physicochemical drivers of community structure, in this study we used community sequencing data to explore the concept of microbial communities as living chemical systems that respond to their environment – that is, geobiochemical systems. We recently described the detection of signals of environmental redox and salinity gradients through compositional analysis of protein sequences coded in shotgun metagenome datasets (*7,8*). Analysis of the more readily available 16S rRNA gene sequencing datasets would enable a more comprehensive test of the hypothesis of a chemical link between communities and environmental conditions. Therefore, in this study, 16S-based community profiles were combined with predicted proteomes from the NCBI RefSeq database to obtain the amino acid compositions of inferred community proteomes, which were then used to compute chemical compositional metrics. In this paper, "predicted proteomes" refers to RefSeq proteomes for particular taxonomic groups, which are automatically predicted from genome sequences, and "inferred community proteomes" refers to the combination of predicted RefSeq proteomes with 16S-based microbial abundance profiles to estimate the amino acid composition of the proteome for the entire community.

We made comparisons at a global scale between reduced hydrothermal environments represented by terrestrial hot springs or submarine vents and seawater, hypersaline lakes, and microbial mats. Remarkably, the inferred community proteomes for hydrothermal environments are shifted toward much lower carbon oxidation state compared to other environments. At a

local scale, oxygen gradients in water columns of the Black Sea, Swiss lakes, Eastern Tropical North Pacific, the Sansha Yongle Blue Hole in the South China Sea, and Ursu Lake in Romania all show a decrease in carbon oxidation state with depth. Moreover, microbial communities in streams affected by unconventional oil and gas (UOG) operations and produced water from UOG wells have lower oxidation state of carbon of the inferred proteomes compared to non-affected streams and the injected hydraulic fracturing fluid. The latter result is consistent with the expectation that fracture networks in organic-rich shales are anoxic environments. By representing shifts in the taxonomic composition of microbial communities as chemical differences, this analysis reveals a possible mechanism for the coupling between microbial abundance patterns and redox gradients in natural and engineered environments.

Results

We used inferred community proteomes to calculate two compositional metrics. Carbon oxidation state (Z_C) of an organic molecule denotes the average charge on carbon atoms required for electroneutrality given formal charges of all other atoms (*9*). Also known by other names such as nominal oxidation state of carbon, Z_C is widely used in studies of natural organic matter, and to an increasing extent for comparing proteins coded by metagenomic sequences (*10, 11*). The equation for Z_C only requires elemental abundances (*8, 12*) and is consistent with models for complex organic matter used, for example, in soil science (*9*).

The second compositional metric, stoichiometric hydration state (n_{H_2O}), is computed from the number of water molecules in the theoretical reaction to form a protein from a set of thermodynamic components, also referred to as basis species. The method for computing n_{H_2O} is specifically derived for proteins by using a set of basis species that reduces the correlation between Z_C and n_{H_2O} for all the proteins coded by a single genome (*Escherichia coli*) in order to more easily visualize independent variations in oxidation state and hydration state (*8*).

This method has been used to separate the signatures of oxidation-reduction and hydration-dehydration reactions in environmental sequences and clinical proteomic data (*8, 13*).

Representative datasets from studies of hydrothermal systems, stratified water bodies, and microbial mats were analyzed to test the predictions that Z_C aligns with redox conditions both locally (within datasets) and globally (across datasets). Accession numbers and literature references for the datasets analyzed here are listed in Table 1. While the main focus of this study is on redox gradients, datasets for a freshwater to marine transition (Baltic Sea) and hypersaline systems were also included in order to assess the prediction that higher salinity exerts a dehydrating force whose effects should be visible in lower n_{H_2O} of inferred community proteomes. To provide an evolutionary context for the analysis of 16S gene sequence datasets, the presentation begins with a visualization of the chemical compositions of the predicted proteomes of taxonomic groups in the RefSeq database.

Chemical differences among taxonomic groups

Compositional metrics for predicted proteomes for phyla with the greatest number of representative lower-level taxa in the RefSeq database are plotted in Fig. 1. These plots show Z_C and n_{H_2O} computed from the mean amino acid composition of RefSeq proteins for phyla and for classes within each phylum. The first panel includes viruses and archaeal and bacterial phyla, and reveals that proteins in many viruses have a lower stoichiometric hydration state than most cellular organisms except for Bacteroidetes. The next panel excludes viruses to consider cellular organisms. The proteomes of organisms affiliated to Bacteroidetes have the lowest overall n_{H_2O}, and those for Crenarchaeota, Fusobacteria, Actinobacteria, and Euryarchaeota (except the classes Haloarchaea and Nanohaloarchaea) have relatively high n_{H_2O}. In terms of Z_C, Actinobacteria, Planctomycetes, and Haloarchaea and Nanohaloarchaea within the Euryarchaeota are the groups with the most oxidized proteomes, whereas Crenarchaeota, Thermotogae, Fu-

sobacteria, and Tenericutes have the most reduced proteomes. The third panel represents chemical composition at a lower taxonomic level, for the proteobacterial classes and their orders. The stoichiometric hydration state distinguishes the main classes of Proteobacteria, with most orders of Alphaproteobacteria and Gammaproteobacteria at high and low n_{H_2O}, respectively, although the between-order variability for Alphaproteobacteria is much higher. Notably, the proteobacterial class with the most reduced proteins is Epsilonproteobacteria. Members of this class are often identified in hydrothermal vent communities (*14, 15*), and in a proposed reclassification this class belongs to a new phylum named Campylobacterota (*16*).

The chemical representation of microbial proteomes generates a plethora of hypotheses about the effects of oxidation-reduction and hydration-dehydration reactions on biochemical evolution. For instance, the highly reducing environmental conditions of sediments and hydrothermal fluids (*17, 18*) provide an explanation based on evolutionary adaptation for the low Z_C of proteins in the Thermococci, Methanococci, and Archaeoglobi, which are the most reduced classes in the Euryarchaeota and among the most reduced of all bacterial classes shown in Fig. 1. In an even broader evolutionary context, the lower hydration state of viral proteomes than those of cellular organisms may not be surprising given the absence of a cytoplasmic compartment in viruses; the tight structure of the viral package is thought to exclude water (*19*). The low n_{H_2O} of Bacteroidetes, which are often enriched in suspended particles in seawater (*20*), aligns with our previous finding of lower n_{H_2O} of proteins coded by metagenomes of particle-associated compared to free-living communities (*8*).

Some natural clustering is evident in Fig. 1, which indicates that the chemical composition of proteomes of classes within a phylum tend to be more similar to each other than to classes in other phyla. An interesting exception is the much higher Z_C and lower n_{H_2O} of Haloarchaea and Nanohaloarchaea than for other euryarchaeotal classes. The lower water activity associated with hypersaline environments may be one reason for these groups to have evolved proteomes

with lower n_{H_2O}. However, the high Z_C of these groups is most likely not an adaptation to more oxidizing conditions; indeed, the solubility of O_2 decreases at higher salinity (*21*). Instead, the proteomes of many halophiles have greater numbers of acidic residues that stabilize the three-dimensional structure of proteins in high-salt conditions; because of the oxygen atoms contained in carboxylic acid groups, this adaptation also results in higher average oxidation state of carbon of the proteins (*8*).

Microbial communities encode redox and salinity gradients

The first set of compositional analyses of inferred community proteomes is shown in Fig. 2. Local redox gradients represented by stratified water bodies including the Black Sea (*22*) and Lake Fryxell (*23*), submarine vents in the Manus Basin (*24*), and within the Guerrero Negro microbial mat (*25*) all provide evidence supporting the predicted change of Z_C. Specifically, Z_C is locally lower in the deep euxinic water of the Black Sea and anoxic water of Lake Fryxell, lower in the hotter water samples for the Manus Basin (which have greater input of reduced hydrothermal fluids), and lower just below the surface of the hypersaline Guerrero Negro microbial mat, which experiences a sharp oxygen gradient during the day (*25*). Mat samples were also taken from the floor of Lake Fryxell, but the environmental gradient here is between relatively oxygenated shallow water and anoxic deeper water (*23*).

No oxygen or redox measurements were reported in the study on alkaline hot spring communities in Yellowstone National Park (*26*); however, this and the Manus Basin dataset represent hydrothermal systems that emit highly reduced fluids. In the index plot in the center of Fig. 2, it can be seen that these datasets are distributed toward lower Z_C than for the lake and seawater environments, thereby supporting the conclusion that Z_C provides a signature of oxidation-reduction conditions on a global scale.

Both the Baltic Sea (*27*) and saline soils in the Qarhan salt lake (*28*) exhibit decreasing n_{H_2O}

with greater salinity. However, the trends for the Qarhan salt lake soils and Tibetan Plateau lake datasets (*29*) are more strongly dominated by increasing Z_C in hypersaline conditions, which is consistent with trends observed from compositional analysis of shotgun metagenomic data in other hypersaline systems (*8*).

Different sample types in some datasets have distinct compositional features. Fauna samples in the Manus Basin dataset have lower n_{H_2O} than the majority of fluid and rock samples. Another interesting result is the higher n_{H_2O} inferred for archaeal proteomes compared to bacterial proteomes in hot springs in Yellowstone National Park. This is aligned with expected phylogenetic differences, since the proteomes of major archaeal groups detected in these samples, including Crenarchaeota and Euryarchaeota (*26*), generally have higher n_{H_2O} than bacterial proteomes (Fig. 1B).

Many water bodies around the world develop vertical redox gradients as a result of microbial respiration of organic matter derived from the photic zone that leads to oxygen depletion with depth. Besides the Black Sea, we analyzed data for permanently stratified lakes in Switzerland (*30*), the oxygen minimum zone of the Eastern Tropical North Pacific (ETNP) (*31*), the Sansha Yongle Blue Hole in the South China Sea (*32*), and Ursu Lake in Central Romania (*33*). At each location, the Z_C of the inferred community proteomes decreases with depth (Fig. 3). At the ETNP, the Z_C decreases strongly with depth in the free-living communities (0.2–1.6 μm size fraction), but to a lesser extent in particle-associated communities (1.6–30 μm size fraction), suggesting that environmental microniches and cell-cell interactions might make these communities less sensitive to external redox conditions.

At a non-stratified location outside the Sansha Yongle Blue Hole, the water remains oxygenated at depth. In the upper 100 meters, the profile of Z_C has a "C" shape, but maintains relatively high values at greater depths. The overall higher Z_C with depth appears to be more closely associated with the ratio of nitrate to nitrate than with O_2 concentration. It therefore ap-

pears that where oxygen is abundant, other inorganic redox couples may reflect redox gradients that could contribute to the environmental shaping of community structure.

Ursu Lake in Central Romania has both strong redox and salinity gradients, and a recent study reported chemical and biological measurements over a span of 9 months to characterize the spatial and seasonal variability (*33*). At all times, both O_2 and Z_C show an overall decline with depth. The profile of Z_C exhibits a broad maximum at less than 2 m depth in November that becomes narrower and deeper through the winter and spring. The development of the Z_C maximum precedes that of an O_2 maximum in February, and the two profiles exhibit a remarkable meter-scale correspondence in April.

Physicochemical signal is strongest at high taxonomic levels

Not only the abundances but also the proteomic composition of individual taxonomic groups contribute to the overall chemical differences between microbial communities. To assess the chemical differences along redox and salinity gradients in more detail, the abundances and calculated Z_C or n_{H_2O} of major taxonomic groups at the domain, phylum, class, and genus levels are plotted in Fig. 4 for the Manus Basin and Baltic Sea. The plots represent aggregated values for the indicated taxonomic groups, but the lowest-level taxonomic classification for each sequence in that group was used for mapping to the NCBI taxonomy to generate the inferred microbial proteome (see Materials and Methods).

The first thing to notice in Fig. 4 is that at the domain level (the leftmost plot), increasing temperature in the Manus Basin is associated with decreased Z_C. The same trend can be seen in Fig. 2; note that the symbol shape and color have the same meaning in both figures. Going down to the phylum level, the high-temperature samples in the Manus Basin are associated with greater numbers of Aquificae and Campylobacterota (formerly Epsilonproteobacteria) and fewer Proteobacteria. These groups have relatively low and high Z_C, respectively, which to a

large extent explains the overall lower Z_C at the whole-community level. However, the campylobacterotal sequences themselves are affiliated with organisms whose proteomes have lower Z_C in the higher-temperature samples. Therefore, the whole-community chemical differences are due to both differential taxonomic abundances at the phylum level, which have the largest effect, as well as differential abundances within phyla. A similar finding applies to the major classes; at this level it is apparent that the proteobacterial contribution is mainly due to lower numbers of Gammaproteobacteria. Notably, the arrows for the two most abundant taxonomic classes (Gammaproteobacteria and Campylobacteria) point in the downward direction, indicating that the taxa identified within these classes have relatively lower Z_C in the higher-temperature samples. The RDP Classifier does not resolve taxonomy below the genus level, so the final plot has horizontal lines; nevertheless, the differential abundances of these genera yield a small Z_C difference between hotter and cooler fluids in the same direction as the whole-community trend.

Analogous reasoning can be used to interpret the trends in the Baltic Sea. The relatively high n_{H_2O} in low-salinity samples is mostly controlled by an increase in Actinobacteria. In contrast, Proteobacteria become less abundant at lower salinity, which to some extent counteracts the n_{H_2O} rise, but the Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria taken individually each exhibit higher n_{H_2O} at lower salinity. The chemical differences among genus-level assignments for the Baltic Sea suggest an opposite trend (higher n_{H_2O} at higher salinity), but this is less likely to represent the actual differences because the low classification rate to the genus (37%; see table S1) together with the 1% abundance cutoff for genera in Fig. 4 results in a low fraction of assignments represented at this level (23%) .

To summarize, differences in the chemical composition of inferred community proteomes along redox and salinity gradients are mainly associated with changes in abundances of particular phyla. Changes at lower taxonomic levels make a smaller contribution but in many cases the chemical differences are in the same direction, indicating that physicochemical conditions

shape microbial communities at multiple taxonomic levels.

Application to datasets for unconventional oil and gas operations

During unconventional oil and gas (UOG) extraction, hydraulic fracturing fluid is injected into shale formations to create extensive horizontal fracture networks that improve hydrocarbon recovery. The injected fluid mixes and reacts with natural formation waters and the fractured rock surfaces. Water that returns to the surface is initially referred to as flowback and later as produced water during the hydrocarbon production stage of the well (*34*). Depending on the operator, chemical oxidants may be added to the injected fracturing fluid; these enhance mineral dissolution and can also have antibacterial effects. Even without such additives, fracturing fluids generally consist of large amounts of water from surface sources, which makes them highly oxidized compared to the reducing conditions in organic-rich shale.

Changes in the chemistry and biology of streams that are in proximity to or may be directly affected by UOG operations are an important issue for environmental assessments of UOG operations. The chemistry of affected surface streams is thought to reflect the possible input of methane and/or chemical additives in fracking fluid from nearby wells, though the overall strength of these associations has been debated (*35*).

Previous authors have noted that little quantitative information is available about the changes in redox conditions of flowback and produced fluid over time (*36, 37*). However, it was found that oxygen is rapidly depleted in flowback and produced waters in the Duvernay Formation in Alberta, Canada (*38*). In a study on the Marcellus Shale in Pennsylvania, USA, the abundances of S-bearing organic compounds determined from FT-ICR-MS measurements exhibited a decrease in carbon oxidation state compared to injected fluids (*39*). Furthermore, oxidation-reduction potential (ORP) measured using a multiprobe is lower (more reducing) in groundwater samples with higher concentrations of CH_4 (*40*). Therefore, we predicted that inferred

community proteomes in produced water would be shifted toward a more reduced state compared to source water. A related prediction is that putative hydrocarbon input to surface streams has a similar reducing effect, but is likely to be much smaller because of the greater extent of dilution.

Analysis of 16S sequence data for water from streams in Northwestern Pennsylvania (*41*) shows that both Z_C and n_{H_2O} are lower in streams affected by Marcellus Shale activity compared to unaffected streams (Fig. 5A). Smaller differences, but in the same direction, are found for stream sediments and for stream water in Pennsylvania State Forests (*35, 42*) (Fig. 5B). Although comparisons between studies are complicated by different sampling strategies and analytical techniques, the results show a general agreement that inferred community proteomes in streams affected by UOG operations are offset toward lower Z_C and n_{H_2O}.

Much larger shifts, toward lower Z_C and higher n_{H_2O}, occur in produced waters compared to source waters and injected fracturing fluids. This trend is evident for not only the Marcellus Shale (*43*) but also the Denver–Julesburg Basin in Colorado, USA (*44*) and the Duvernay Formation (*38*) (Fig. 5, C and D).

Discussion

The main finding of this study is that the oxidation state of inferred community proteomes decreases in more reducing conditions at global and local scales. This conclusion is consistent with our earlier analysis of shotgun metagenomic data that showed lower Z_C of proteins in Yellowstone hot springs compared to other environments (*7*) and a study from another group reporting that metagenome-encoded proteins for the newly discovered Old City hydrothermal field have lower Z_C than ambient seawater (*11*) (the value for Z_C of seawater communities used by (*11*) was taken from our previous study (*7*)). The present study also reveals decreasing Z_C with depth using data from multiple studies for stratified water bodies; this strong physicochem-

ical signal is in contrast to the ambiguous results for oxygen minimum zones that we previously obtained by analyzing shotgun metagenomic data (*7*).

The results for the Baltic Sea and Qarhan Salt Lake support the prediction that n_{H_2O} decreases in more saline environments (*8*), but in hypersaline environments the compositional trend is mainly toward higher Z_C instead of lower n_{H_2O}. This is again consistent with our previous findings based on shotgun metagenome sequences (*8*), and indicates a limitation of using n_{H_2O} as an indicator of hypersaline conditions.

The ability to use community profiles to directly predict physicochemical associations is a powerful new tool for microbial ecology and can be applied to other systems that are dominated by redox gradients. In UOG systems, the mixing of fracturing fluids with highly saline and anoxic formation waters leads to environmental filtering for halophilic and anaerobic organisms (*36*). Produced waters from many shales converge toward a common profile dominated by *Halanaerobium* (*36*), but in the Denver–Julesburg Basin *Thermoanaerobacter*, which has similar metabolic capabilities, is present instead (*44*). These groups are both members of the phylum Clostridia, which has representatives with more reduced proteomes than many other bacterial classes (Fig. 1B). Specifically, the predicted RefSeq proteomes of *Halanaerobium* and *Thermoanaerobacter* have Z_C values of -0.195 and -0.227, respectively (see file `RefSeq_metrics.csv` in the JMDplots package); the very low oxidation states of these abundant groups accounts for much of the decrease in Z_C in produced fluids. This trend is the opposite of increasing Z_C observed in other hypersaline systems (this study and (*8*)), which strengthens the interpretation that reducing conditions are a primary driver of community structure in UOG produced water. Z_C decreases to a lesser extent in UOG-affected streams (Fig. 5B), but the consistency among datasets suggests a new line of evidence that reducing conditions may contribute to shape the community ecology in these systems.

Under a mass-action hypothesis, higher salinity should have a dehydrating effect, but this

prediction is falsified by the observed increase of n_{H_2O} of inferred community proteomes in produced fluids. It is possible that lower Z_C is intrinsically linked to higher n_{H_2O} as a result of the background correlation between these metrics when n_{H_2O} is calculated using the QEC basis species (Fig. 5C; see also (*8*)). However, datasets for the stream samples show the opposite trend, so the relations between Z_C and n_{H_2O} reflect specific communities and are not universally dictated by the basis species. More work is needed to derive a chemical metric that better captures the relationship between the chemical composition of community proteomes and salinity, which is a strong driver of microbial community structure (*6*). At the same time, Lozupone and Knight (2007) (*6*) remarked that "oxygenation may also be important" for the global distribution of bacteria, such as the presence of anaerobic Clostridia in sediments. The present analysis shows that to a considerable extent the association between particular taxa and oxygen concentrations can be predicted from differences of Z_C.

The results of the analysis applied to UOG systems suggest that in addition to salinity, the large redox gradient between the surface (oxidizing) and subsurface (reducing) is a primary factor that shapes the structure of microbial communities. In contrast, a dependence on redox conditions was not identified in the multivariate analyses previously reported for these systems (*38, 43, 44*). This may be because ordination methods are limited to the available chemical measurements. Previous authors have commented on the dearth of redox and oxygen measurements in samples collected from black shale well sites (*36*), and a lack of such measurements is also apparent by analyzing metadata for the USGS National Produced Waters Geochemical Database (*45*). Of the 114943 records in the database, there are only 66 with measurements of dissolved oxygen (O_2) or ORP, and these are only for conventional or geothermal wells. Since oxygen or redox measurements of produced waters from unconventional wells were not available, the multivariate analyses used in previous studies could not be used to identify an association between microbial communities and oxidation-reduction state of the fluids. However,

such an association is exactly what is predicted by the compositional analysis in this study.

In spite of the oxygenation of injected fluids, subsurface conditions are likely to be anoxic (*38*), so other types of measurements should be considered for monitoring the redox state of produced water. For instance, the USGS database cites a study from 2009 that gives both NO_3^- and NO_2^- measurements in injected and produced water from the Marcellus Shale (*46*). Monitoring the dynamics of N species in conjunction with biological sampling would give further insight into the extent of nitrate reduction in the subsurface (*43, 47*) and could also serve as a proxy for *in situ* redox conditions through a metric such as the NO_3^-/NO_2^- ratio (*48*). Because the electroactivity of specific redox couples is affected by kinetic barriers that are not well understood, ORP measurements generally have a more difficult interpretation. Nevertheless, they can yield information about electrochemical reactions that are relevant to microbial growth, especially if the measurements are made continuously in time (*49*).

By leveraging the chemical information contained in protein sequences, it is possible to achieve a broader view of the coupling between inorganic and organic oxidation-reduction reactions that is essential for all ecosystems (*47, 50*). How geochemistry affects biochemistry is one of the questions addressed in the emerging field of geobiochemistry (*51*), and the present analysis demonstrates a strong linkage between environmental conditions and the chemical composition of inferred community proteomes. Unlike most multivariate techniques used to analyze microbial abundance data, in which the plot axes represent dataset-dependent synthetic variables that can be difficult to interpret (*2*), a chemical representation of communities uses variables that have an intuitive meaning and stable definition, which enables comparisons across datasets. The results suggest that more comprehensive monitoring of dissolved oxygen concentrations and other redox indicators should be used to better characterize the responses of microbial communities in produced water and streams affected by unconventional hydrocarbon extraction and ecosystems in general.

Materials and Methods

Data sources, processing, and classification

16S rRNA gene sequences were downloaded from the NCBI Sequence Read Archive (SRA) except for the Guerrero Negro microbial mat sequences (*25*), which were obtained from GenBank. The processing pipeline consisted of merging of paired-end reads, length and quality filtering, removal of singletons, subsampling, chimera removal, and taxonomic classification. VSEARCH version 2.15.0 (*52*) was used to merge Illumina paired-end reads; for some datasets with low-quality reverse reads, only the forward reads were used as in previous studies (*41, 42*). For Illumina datasets, quality and length filtering were done with the options `-fastq_maxee_rate 0.005` (i.e. maximum one expected error for every 200 bases) and `-fastq_trunclen` with a length value depending on the specific dataset. For 454 datasets, where reads are generally longer but have more variable length, quality and length filtering were done with `-fastq_truncqual 15 -fastq_minlen 200 -fastq_maxlen 600`; also, the option `-fastq_stripleft 18` was used to remove adapter sequences. Sequence processing statistics and additional details are given in table S1.

After filtering, the remaining sequences for all samples in each dataset were pooled and singletons (sequences that appear exactly once, but not including subsequences of other sequences) were removed. Then, samples were subsampled to a depth of 10000 sequences; samples with fewer than 10000 sequences were not affected. The subsampling reduces the processing time for chimera detection, which is the longest step in the pipeline, and retains enough sequences for classifying the major taxonomic groups in the communities. Reference-based chimera detection was performed using the VSEARCH command `-uchime_ref` with the SILVA 138.1 SSURef NR99 database (*53*). Sequences identified as chimeras or borderline chimeras were removed.

The remaining sequences were processed with the Ribosomal Database Project (RDP) (*54*) Classifier version 2.13 (*55*) with the provided training set (RDP 16S rRNA training set No. 18 07/2020). The sequence classifications for all samples in each dataset were merged using the RDP Classifier command `merge-count`.

RefSeq proteomes

Predicted protein sequences were obtained for all 49448 bacterial, archaeal and viral taxa in the RefSeq database release 206 (2020-05-21) (*56*). For each taxon (identified by a unique taxid number), the amino acid compositions of all protein sequences were summed to generate the total amino acid composition of the predicted proteome. For each taxid, the available taxonomic names at ranks of superkingdom, phylum, class, order, family, genus, and species were parsed from the current (as of the RefSeq release date) NCBI taxonomy files.

For taxonomic groups at genus and higher levels, the amino acid compositions of proteins present in all taxa within this group (including lower levels) were combined and normalized to yield the overall amino acid composition of the predicted proteome for this group. The number of proteomes, corresponding to the number of taxonomic groups at each level, is 4788 (genus), 763 (family), 303 (order), 140 (class), 78 (phylum), and 3 (superkingdom).

Taxonomy mapping

To infer the amino acid compositions of communities from 16S sequencing data, RDP classifications at only the root or domain level were first omitted because they provide very little taxonomic resolution. Sequences assigned to RDP class- and family-level name Chloroplast or genus-level names Chlorophyta and Bacillariophyta were also discarded because they do not fall within the archaeal and bacterial taxonomy used by NCBI. All remaining classifications at any taxonomic level were retained for attempted mapping to the NCBI taxonomy.

In general, mapping from the RDP Classifier to the NCBI taxonomy was performed by text matching of both the taxonomic rank and name. Some particular mappings were used to improve the representation of common taxa in the datasets. The RDP phylum Cyanobacteria/Chloroplast, class Planctomycetacia, and genus *Escherichia/Shigella* were mapped to the NCBI phylum Cyanobacteria, class Planctomycetia, and genus *Escherichia*, respectively. The RDP order Rhizobiales was mapped to the NCBI order Hyphomicrobiales (*57*). The RDP taxon Spartobacteria genera *incertae sedis*, which is relatively abundant in the Baltic Sea (*27*), was mapped to NCBI class Spartobacteria. The RDP taxon Marinimicrobia genera *incertae sedis*, which was identified in this study in some deep ocean datasets (*24, 31*), was mapped to NCBI species Candidatus Marinimicrobia bacterium, which is the only representative of the Candidatus Marinimicrobia phylum in the RefSeq database. Among Acidobacteria, which are fairly abundant in river water and sediment, the RDP genus-level classifications Gp1 and Gp6 were mapped to NCBI genera *Acidobacterium* and *Luteitalea*, respectively, which are members of Acidobacteria subdivisions 1 and 6 (*58*). The RDP genus-level cyanobacterial groups GpI, GpIIa, and GpVI were mapped to the NCBI genera *Nostoc*, *Synechococcus*, and *Pseudanabaena*, and the RDP taxon Family II was mapped to the family Synechococcaceae; these mappings are based on names of members of these groups given in Bergey's Manual (*59*), although the mappings are necessarily imperfect because of inconsistencies with the NCBI taxonomy.

Any other RDP classifications whose rank and name could not be matched to the NCBI taxonomy were removed from the subsequent calculations. Across all datasets, a median of 95.5% of RDP classifications at all levels from genus to phylum were mapped to the NCBI taxonomy (table S1). The lowest classification rate is for a dataset for UOG produced water (*44*), in which the genus *Cavicella* makes up 27% of the RDP classifications but has no counterpart in the available RefSeq proteomes (table S2).

Compositional analysis

The RDP counts for each mapped taxon were multiplied by the amino acid compositions of the RefSeq taxa described above and summed to obtain the amino acid composition of the inferred whole community proteome. Only unique classifications at the lowest taxonomic level were included in the sum. Therefore, the inferred community proteome represents counts of all sequences classified and mapped at the genus level together with counts of sequences classified and mapped at each higher level, up to phylum, for which lower-level assignments were not generated by the RDP Classifier. The `ZCAA()` and `H2OAA()` functions in the canprot package (*13*) were used to calculate Z_C and n_{H_2O} from the amino acid compositions.

References

1. P. D. Schloss, S. L. Westcott, Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **77**, 3219–3226 (2011).

2. A. Ramette, Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62**, 142–160 (2007).

3. J. Wang, J. Shen, Y. Wu, C. Tu, J. Soininen, J. C. Stegen, J. He, X. Liu, L. Zhang, E. Zhang, Phylogenetic beta diversity in bacterial assemblages across ecosystems: Deterministic versus stochastic processes. *ISME J.* **7**, 1310–1321 (2013).

4. L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vázquez-Baeza, A. González, J. T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N. A. Bokulich, J. Lefler, C. J. Brislawn,

G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, R. Knight, The Earth Microbiome Project Consortium, A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

5. E. B. Alsop, E. S. Boyd, J. Raymond, Merging metagenomics and geochemistry reveals environmental controls on biological diversity and evolution. *BMC Ecol.* **14**, 16 (2014).

6. C. A. Lozupone, R. Knight, Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci.* **104**, 11436–11440 (2007).

7. J. M. Dick, M. Yu, J. Tan, A. Lu, Changes in carbon oxidation state of metagenomes along geochemical redox gradients. *Front. Microbiol.* **10**, 120 (2019).

8. J. M. Dick, M. Yu, J. Tan, Uncovering chemical signatures of salinity gradients through compositional analysis of protein sequences. *Biogeosciences* **17**, 6145–6162 (2020).

9. F. Visconti, J. M. de Paz, Estimation of the carbon valence from its average formal oxidation state in the soil organic matter. *Eur. J. Soil Sci.* **n/a** (2021). `https://doi.org/10.1111/ejss.13122`.

10. E. M. Fones, D. R. Colman, E. A. Kraus, D. B. Nothaft, S. Poudel, K. R. Rempfert, J. R. Spear, A. S. Templeton, E. S. Boyd, Physiological adaptations to serpentinization in the Samail Ophiolite, Oman. *ISME J.* **13**, 1750–1762 (2019).

11. A. Lecoeuvre, B. Ménez, M. Cannat, V. Chavagnac, E. Gérard, Microbial ecology of the newly discovered serpentinite-hosted Old City hydrothermal field (southwest Indian ridge). *ISME J.* **15**, 818–832 (2021).

12. J. M. Dick, Average oxidation state of carbon in proteins. *J. R. Soc. Interface* **11**, 20131095 (2014).

13. J. M. Dick, Water as a reactant in the differential expression of proteins in cancer. *Comput. Syst. Oncol.* **1**, e1007 (2021).

14. S. Nakagawa, K. Takai, F. Inagaki, H. Hirayama, T. Nunoura, K. Horikoshi, Y. Sako, Distribution, phylogenetic diversity and physiological characteristics of epsilon-*Proteobacteria* in a deep-sea hydrothermal field. *Environ. Microbiol.* **7**, 1619–1632 (2005).

15. B. J. Campbell, A. S. Engel, M. L. Porter, K. Takai, The versatile ϵ-proteobacteria: Key players in sulphidic habitats. *Nat. Rev. Microbiol.* **4**, 458–468 (2006).

16. D. W. Waite, I. Vanwonterghem, C. Rinke, D. H. Parks, Y. Zhang, K. Takai, S. M. Sievert, J. Simon, B. J. Campbell, T. E. Hanson, T. Woyke, M. G. Klotz, P. Hugenholtz, Addendum: Comparative genomic analysis of the class *Epsilonproteobacteria* and proposed reclassification to Epsilonbacteraeota (phyl. nov.). *Front. Microbiol.* **9**, 772 (2018).

17. D. R. Lovley, M. J. Klug, Model for the distribution of sulfate reduction and methanogenesis in freshwater sediments. *Geochim. Cosmochim. Acta* **50**, 11–18 (1986).

18. T. J. Lin, H. C. Ver Eecke, E. A. Breves, M. D. Dyar, J. W. Jamieson, M. D. Hannington, H. Dahle, J. L. Bishop, M. D. Lane, D. A. Butterfield, D. S. Kelley, M. D. Lilley, J. A. Baross, J. F. Holden, Linkages between mineralogy, fluid chemistry, and microbial communities within hydrothermal chimneys from the Endeavour Segment, Juan de Fuca Ridge. *Geochem. Geophys. Geosyst.* **17**, 300–323 (2016).

19. F. H. Westheimer, Why nature chose phosphates. *Science* **235**, 1173–1178 (1987).

20. B. Fernández-Gómez, M. Richter, M. Schüler, J. Pinhassi, S. G. Acinas, J. M. González, C. Pedrós-Alió, Ecology of marine Bacteroidetes: A comparative genomics approach. *ISME J.* **7**, 1026–1037 (2013).

21. J. E. Sherwood, F. Stagnitti, M. J. Kokkinn, W. D. Williams, Dissolved oxygen concentrations in hypersaline waters. *Limnol. Oceanogr.* **36**, 235–250 (1991).

22. M. Sollai, L. Villanueva, E. C. Hopmans, G.-J. Reichart, J. S. Sinninghe Damsté, A combined lipidomic and 16S rRNA gene amplicon sequencing approach reveals archaeal sources of intact polar lipids in the stratified Black Sea water column. *Geobiology* **17**, 91–109 (2019).

23. A. D. Jungblut, I. Hawes, T. J. Mackey, M. Krusor, P. T. Doran, D. Y. Sumner, J. A. Eisen, C. Hillman, A. K. Goroncy, Microbial mat communities along an oxygen gradient in a perennially ice-covered Antarctic lake. *Appl. Environ. Microbiol.* **82**, 620–630 (2016).

24. D. V. Meier, P. Pjevac, W. Bach, S. Hourdez, P. R. Girguis, C. Vidoudez, R. Amann, A. Meyerdierks, Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J.* **11**, 1545–1558 (2017).

25. J. K. Harris, J. G. Caporaso, J. J. Walker, J. R. Spear, N. J. Gold, C. E. Robertson, P. Hugenholtz, J. Goodrich, D. McDonald, D. Knights, P. Marshall, H. Tufo, R. Knight, N. R. Pace, Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME J.* **7**, 50–60 (2013).

26. K. Bowen De León, R. Gerlach, B. M. Peyton, M. W. Fields, Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Front. Microbiol.* **4**, 330 (2013).

27. D. P. R. Herlemann, D. Lundin, A. F. Andersson, M. Labrenz, K. Jürgens, Phylogenetic signals of salinity and season in bacterial community composition across the salinity gradient of the Baltic Sea. *Front. Microbiol.* **7**, 1883 (2016).

28. K. H. Xie, Y. Deng, S. C. Zhang, W. H. Zhang, J. R. Liu, Y. L. Xie, X. Z. Zhang, H. Huang, Prokaryotic community distribution along an ecological gradient of salinity in surface and subsurface saline soils. *Sci. Rep.* **7**, 13332 (2017).

29. Z.-P. Zhong, Y. Liu, L.-L. Miao, F. Wang, L.-M. Chu, J.-L. Wang, Z.-P. Liu, Prokaryotic community structure driven by salinity and ionic concentrations in plateau lakes of the Tibetan Plateau. *Appl. Environ. Microbiol.* **82**, 1846–1858 (2016).

30. M. J. Mayr, M. Zimmermann, C. Guggenheim, A. Brand, H. Bürgmann, Niche partitioning of methane-oxidizing bacteria along the oxygen-methane counter gradient of stratified lakes. *ISME J.* **14**, 274–287 (2020).

31. S. Ganesh, L. A. Bristow, M. Larsen, N. Sarode, B. Thamdrup, F. J. Stewart, Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *ISME J.* **9**, 2682–2696 (2015).

32. P. He, L. Xie, X. Zhang, J. Li, X. Lin, X. Pu, C. Yuan, Z. Tian, J. Li, Microbial diversity and metabolic potential in the stratified Sansha Yongle Blue Hole in the South China Sea. *Sci. Rep.* **10**, 5949 (2020).

33. A. Baricz, C. M. Chiriac, A.-Ş. Andrei, P.-A. Bulzu, E. A. Levei, O. Cadar, K. P. Battes, M. Cîmpean, M. Şenilă, A. Cristea, V. Muntean, M. Alexe, C. Coman, E. K. Szekeres, C. I. Sicora, A. Ionescu, D. Blain, W. K. O'Neill, J. Edwards, J. E. Hallsworth, H. L. Banciu, Spatio-temporal insights into microbiology of the freshwater-to-hypersaline, oxic-hypoxic-

euxinic waters of Ursu Lake. *Environ. Microbiol.* **n/a** (2020). https://doi.org/10.1111/1462-2920.14909.

34. H. J. Khan, E. Spielman-Sun, A. D. Jew, J. Bargar, A. Kovscek, J. L. Druhan, A critical review of the physicochemical impacts of water chemistry on shale in hydraulic fracturing systems. *Environ. Sci. Technol.* **55**, 1377–1394 (2021).

35. A. C. Mumford, K. O. Maloney, D. M. Akob, S. Nettemann, A. Proctor, J. Ditty, L. Ulsamer, J. Lookenbill, I. M. Cozzarelli, Shale gas development has limited effects on stream biology and geochemistry in a gradient-based, multiparameter study in Pennsylvania. *Proc. Natl. Acad. Sci.* **117**, 3670–3677 (2020).

36. P. J. Mouser, M. Borton, T. H. Darrah, A. Hartsock, K. C. Wrighton, Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol. Ecol.* **92**, fiw166 (2016).

37. T. Liden, I. C. Santos, Z. L. Hildenbrand, K. A. Schug, *Evaluating Water Quality to Prevent Future Disasters*, S. Ahuja, ed. (Academic Press, 2019), vol. 11 of *Separation Science and Technology*, pp. 199–217.

38. C. Zhong, J. Li, S. L. Flynn, C. L. Nesbø, C. Sun, K. von Gunten, B. D. Lanoil, G. G. Goss, J. W. Martin, D. S. Alessi, Temporal changes in microbial community composition and geochemistry in flowback and produced water from the Duvernay formation. *ACS Earth Space Chem.* **3**, 1047–1057 (2019).

39. J. L. Luek, M. Harir, P. Schmitt-Kopplin, P. J. Mouser, M. Gonsior, Organic sulfur fingerprint indicates continued injection fluid signature 10 months after hydraulic fracturing. *Environ. Sci.: Processes Impacts* **21**, 206–213 (2019).

40. S. T. M. LeDoux, A. Szynkiewicz, A. M. Faiia, M. A. Mayes, M. L. McKinney, W. G. Dean, Chemical and isotope compositions of shallow groundwater in areas impacted by hydraulic fracturing and surface mining in the Central Appalachian Basin, Eastern United States. *Appl. Geochem.* **71**, 73–85 (2016).

41. N. Ulrich, V. Kirchner, R. Drucker, J. R. Wright, C. J. McLimans, T. C. Hazen, M. F. Campa, C. J. Grant, R. Lamendella, Response of aquatic bacterial communities to hydraulic fracturing in northwestern Pennsylvania: A five-year study. *Sci. Rep.* **8**, 5683 (2018).

42. J. R. Chen See, N. Ulrich, H. Nwanosike, C. J. McLimans, V. Tokarev, J. R. Wright, M. F. Campa, C. J. Grant, T. C. Hazen, J. M. Niles, D. Ressler, R. Lamendella, Bacterial biomarkers of Marcellus Shale activity in Pennsylvania. *Front. Microbiol.* **9**, 1697 (2018).

43. M. A. Cluff, A. Hartsock, J. D. MacRae, K. Carter, P. J. Mouser, Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus Shale gas wells. *Environ. Sci. Technol.* **48**, 6508–6517 (2014).

44. N. M. Hull, J. S. Rosenblum, C. E. Robertson, J. K. Harris, K. G. Linden, Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. *Sci. Total Environ.* **644**, 183–192 (2018).

45. M. S. Blondes, K. D. Gans, M. A. Engle, Y. K. Kharaka, M. E. Reidy, V. Saraswathula, J. Thordsen, E. L. Rowan, E. A. Morrissey, U.S. Geological Survey National Produced Waters Geochemical Database (ver. 2.3, January 2018), *Data Release*, U. S. Geological Survey (2018). `https://doi.org/10.5066/F7J964W8`.

46. T. Hayes, Sampling and Analysis of Water Streams Associated with the Development of Marcellus Shale Gas, *Final Report*, Marcellus Shale Coalition

(2009). https://www.water-research.net/naturalgasPA/pdffiles/MSCommission-Report.pdf Accessed on 2021-05-27.

47. B. N. Orcutt, J. B. Sylvan, N. J. Knab, K. J. Edwards, Microbial ecology of the dark ocean above, at, and below the seafloor. *Microbiol. Mol. Biol. Rev.* **75**, 361–422 (2011).

48. A.-L. Ducluzeau, B. Schoepp-Cothenet, R. van Lis, F. Baymann, M. J. Russell, W. Nitschke, The evolution of respiratory O_2/NO reductases: An out-of-the-phylogenetic-box perspective. *J. R. Soc. Interface* **11**, 20140196 (2014).

49. E. Markelova, C. T. Parsons, R.-M. Couture, C. M. Smeaton, B. Madé, L. Charlet, P. Van Cappellen, Deconstructing the redox cascade: What role do microbial exudates (flavins) play? *Environ. Chem.* **14**, 515–524 (2017).

50. A. J. Burgin, W. H. Yang, S. K. Hamilton, W. L. Silver, Beyond carbon and nitrogen: How the microbial energy economy couples elemental cycles in diverse ecosystems. *Front. Ecol. Environ.* **9**, 44–52 (2011).

51. E. L. Shock, E. S. Boyd, Principles of geobiochemistry. *Elements* **11**, 395–401 (2015).

52. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

53. C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F. O. Glöckner, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).

54. J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, J. M. Tiedje, Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2013).

55. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

56. N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, K. D. Pruitt, Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

57. A. Hördt, M. G. López, J. P. Meier-Kolthoff, M. Schleuning, L.-M. Weinhold, B. J. Tindall, S. Gronow, N. C. Kyrpides, T. Woyke, M. Göker, Analysis of 1,000+ type-strain genomes substantially improves taxonomic classification of Alphaproteobacteria. *Front. Microbiol.* **11**, 468 (2020).

58. S. A. Eichorst, D. Trojan, S. Roux, C. Herbold, T. Rattei, D. Woebken, Genomic insights into the *Acidobacteria* reveal strategies for their success in terrestrial environments. *Environ. Microbiol.* **20**, 1041–1063 (2018).

59. D. R. Boone, R. W. Castenholz, G. M. Garrity, eds., *Bergey's Manual® of Systematic Bacteriology* (Springer, 2001).

60. J. M. Dick, JMDplots 1.2.6, Zenodo https://doi.org/10.5281/zenodo.4779010 (2021).

61. R Core Team, *R: A Language and Environment for Statistical Computing (Version 4.1.0)*, R Foundation for Statistical Computing, Vienna, Austria (2021).

62. L. Xie, B. Wang, X. Pu, M. Xin, P. He, C. Li, Q. Wei, X. Zhang, T. Li, Hydrochemical properties and chemocline of the Sansha Yongle Blue Hole in the South China Sea. *Sci. Total Environ.* **649**, 1281–1292 (2019).

63. Q. Wang, RDP Classifier 2.13 (July 2020) Released, *Release Notes*, Center for Microbial Ecology, Michigan State University (2000). https://sourceforge.net/p/rdp-classifier/news/2020/07/rdp-classifier-213-july-2020-release-note/ Accessed on 2021-06-01.

64. E. Paradis, K. Schliep, ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

Acknowledgments

Funding: This research was supported by the National Natural Science Foundation of China (grant nos. 72088101 and 41872151). **Author contributions:** J.D.: Conceptualization, Software, Formal analysis, Writing – original draft preparation; J.T.: Conceptualization, Funding acquisition. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The JMDplots package version 1.2.6 deposited on Zenodo with accession number 4779010 (*60*) has the scripts used for this study and processed data files including amino acid compositions computed from RefSeq (in the directory `extdata/refseq`) and RDP Classifier results (in the directory `extdata/comp16S/RDP`).

A later version of the package was used to make the figures in this submission and will be deposited on Zenodo before final publication (`https://github.com/jedick/JMDplots`). All figures were made using R (*61*) with data files and code provided in the JMDplots package; the "geo16S" vignette in the package runs the functions to make each of the figures.

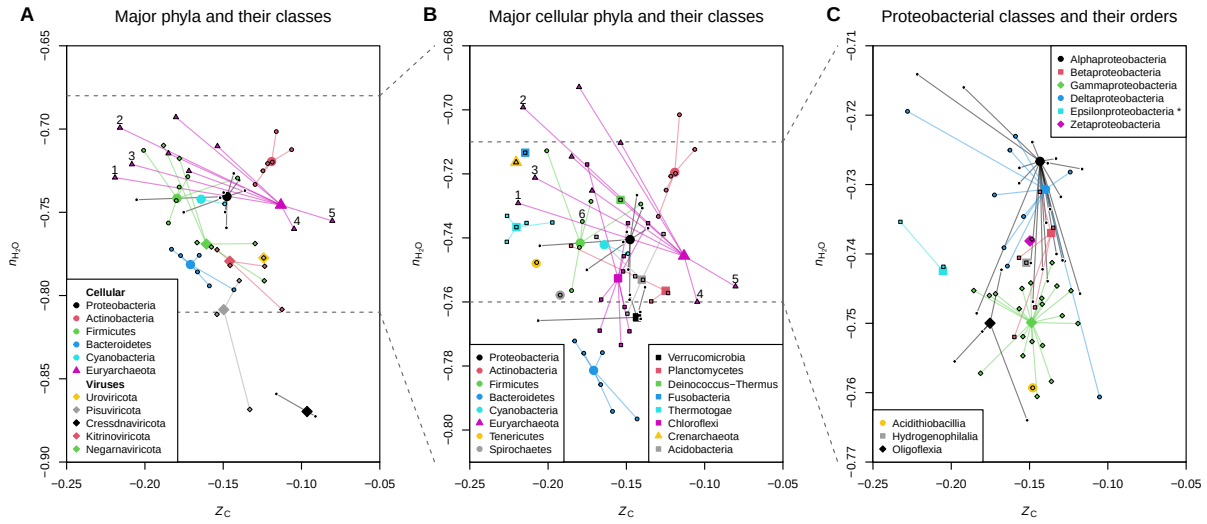

Figure 1: Distinct chemical compositions of predicted proteomes of major taxonomic groups.

Stoichiometric hydration state (n_{H_2O}) and carbon oxidation state (Z_C) are aggregate values for all available proteins for particular taxonomic group in the RefSeq database. (**A**) Archaeal, bacterial, and viral phyla with more than 500 members at all lower levels (identified by unique taxids in the NCBI taxonomy database); (**B**) archaeal and bacterial phyla with more than 60 lower-level members; (**C**) proteobacterial classes. Large symbols are for high-level taxa (phyla in **A** and **B**; classes in **C**) and small outlined symbols represent lower-level taxa (classes in **A** and **B**; families in **C**). Points labeled 1, 2, 3, 4, and 5 in (**A**) and (**B**) are for the euryarchaeotal classes Thermococci, Methanococci, Archaeoglobi, Nanohaloarchaea, and Halobacteria, respectively, and the point labeled 6 in (**B**) is for the class Clostridia in the phylum Firmicutes. The taxonomic names are taken from the current NCBI taxonomy, but according to a proposed reclassification the Epsilonproteobacteria are moved to the Campylobacterota (phyl. nov.) (*16*).

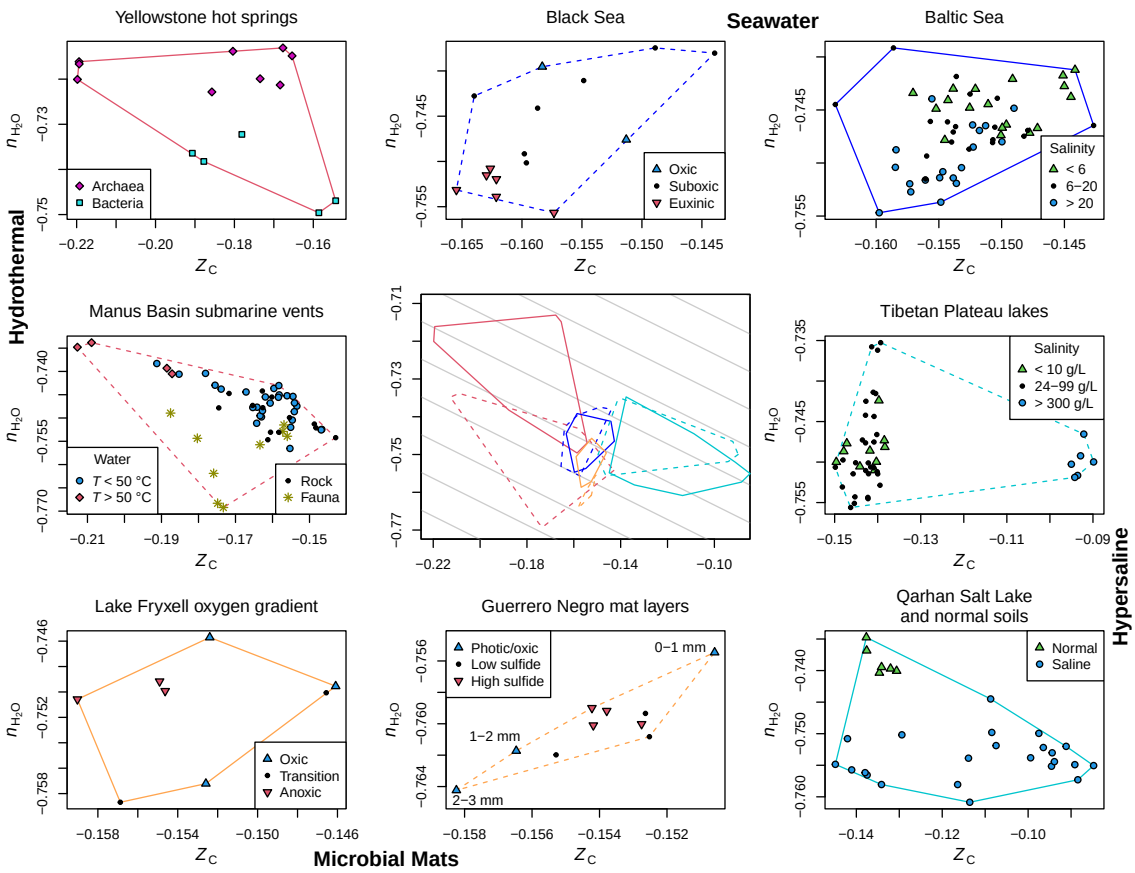

Figure 2: Inferred community proteomes from different environments have distinct chemical signatures.

n_{H_2O} and Z_C were calculated for inferred community proteomes using 16S rRNA gene sequencing datasets for hydrothermal systems, seawater, hypersaline environments, and microbial mats. Sources of data are listed in Table 1. The plot for each dataset shows individual samples as points and the convex hull containing all the samples. The convex hulls for individual datasets are assembled in the center index plot. The gray lines here have a slope corresponding to that of the regression between n_{H_2O} and Z_C for amino acids, and therefore represent the background covariation between these metrics when n_{H_2O} is calculated using the QEC basis species (glutamine, glutamic acid, cysteine, H_2O, and O_2) (*8*). The salinity values reported for the Baltic Sea (supplementary information of (*27*)) have no units.

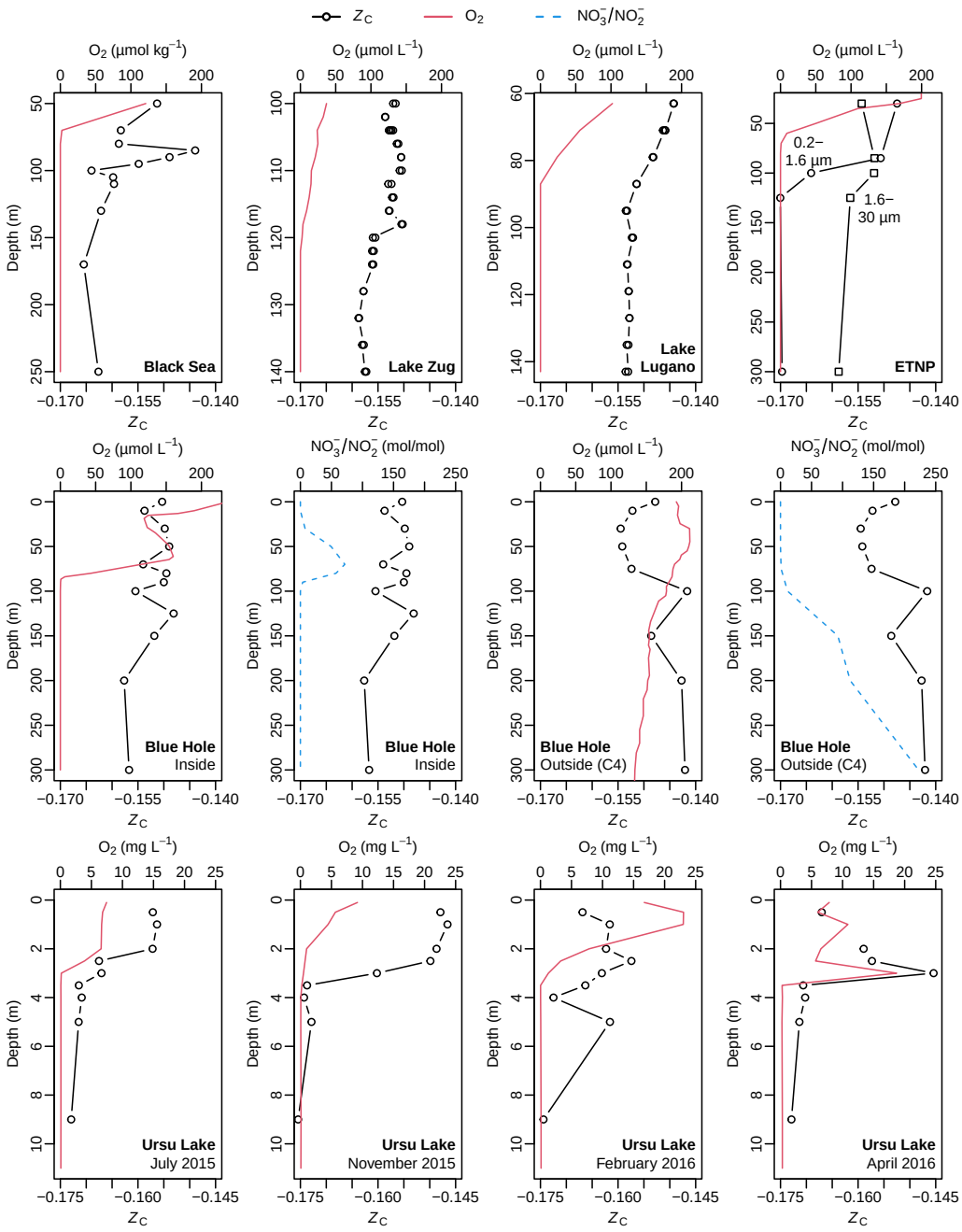

Figure 3: Lower carbon oxidation state is tied to oxygen depletion in water columns.

Depth profiles of O_2 concentrations in water bodies and Z_C of community proteomes inferred from microbial 16S rRNA gene sequences. All sites except for ETNP and station C4 outside the Blue Hole are permanently stratified. Oxygen concentrations were taken from the source publications (see Table 1), except for locations inside and outside the Blue Hole (*62*). For the Blue Hole, ratios of nitrate to nitrite (NO_3^- / NO_2^-) are also plotted based on NO_3^- and NO_2^- concentrations reported in (*62*). No Z_C value is shown for 1 m depth in Ursu Lake in April 2016 because only 161 sequences remained for this sample after all sequence processing steps.

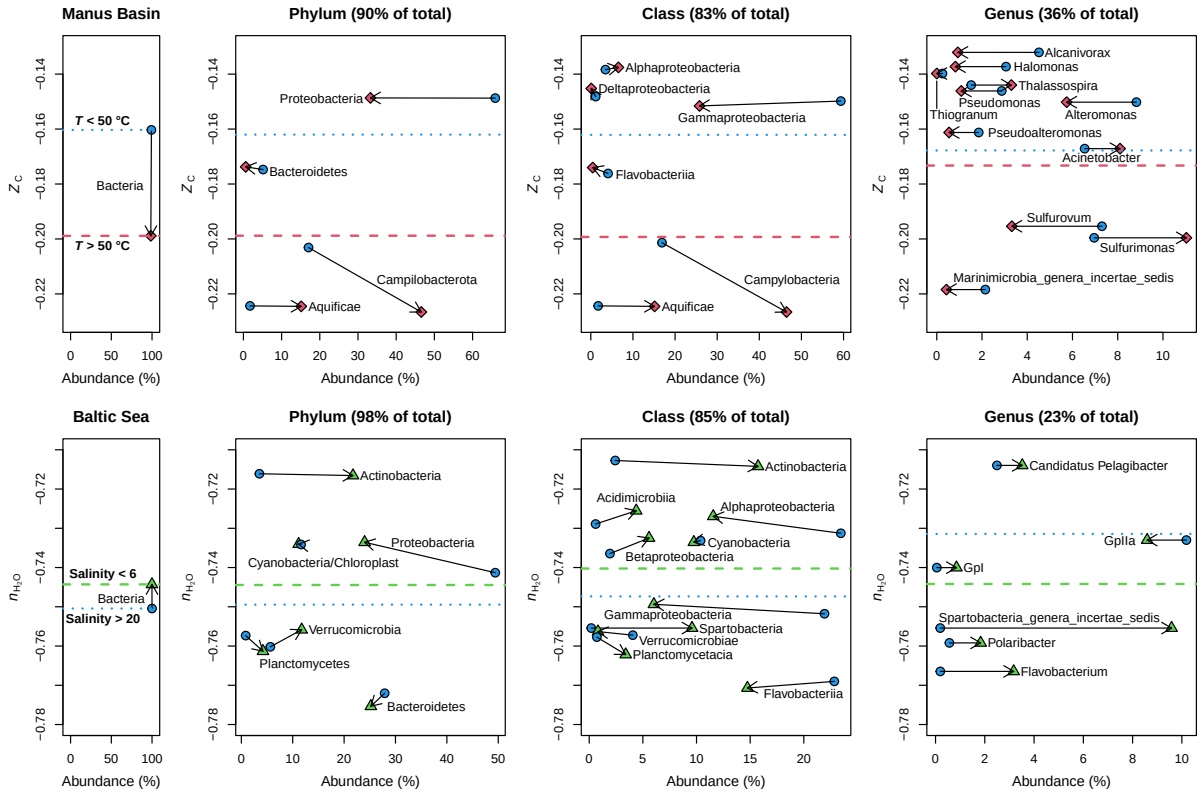

Figure 4: Changes of abundance and chemical composition for individual taxonomic groups.

Symbols represent median values of Z_C for samples with relatively high and low temperature (as a proxy for reducing and oxidizing conditions) from the Manus Basin or n_{H_2O} for samples with high and low salinity from the Baltic Sea. The leftmost plots represent all sequences classified by the RDP Classifier and mapped to the NCBI taxonomy at the domain level; only bacterial sequences are available in these datasets. In subsequent plots, sequences classified and mapped at lower taxonomic levels were combined to calculate the percentage abundance and median Z_C or n_{H_2O} for each group within that level whose sequences make up at least 2% (for phylum or class) or 1% (for genus) of the total number of sequences from all samples. Percentages in the plot titles indicate the total percentage represented by groups shown in the plot. Arrows connect the same taxonomic group in different sample groups and point to samples with higher temperature (Manus Basin) or lower salinity (Baltic Sea). Abundance-weighted means for the taxonomic groups shown in each plot are indicated by dashed lines for high temperature or low salinity samples and dotted lines for low temperature or high salinity samples. All taxonomic names are taken from the output of RDP Classifier. Campilobacterota may be a misspelling of the phylum name Campylobacterota (*16*); the latter spelling is used in the RDP Classifier Release Notes (*63*).

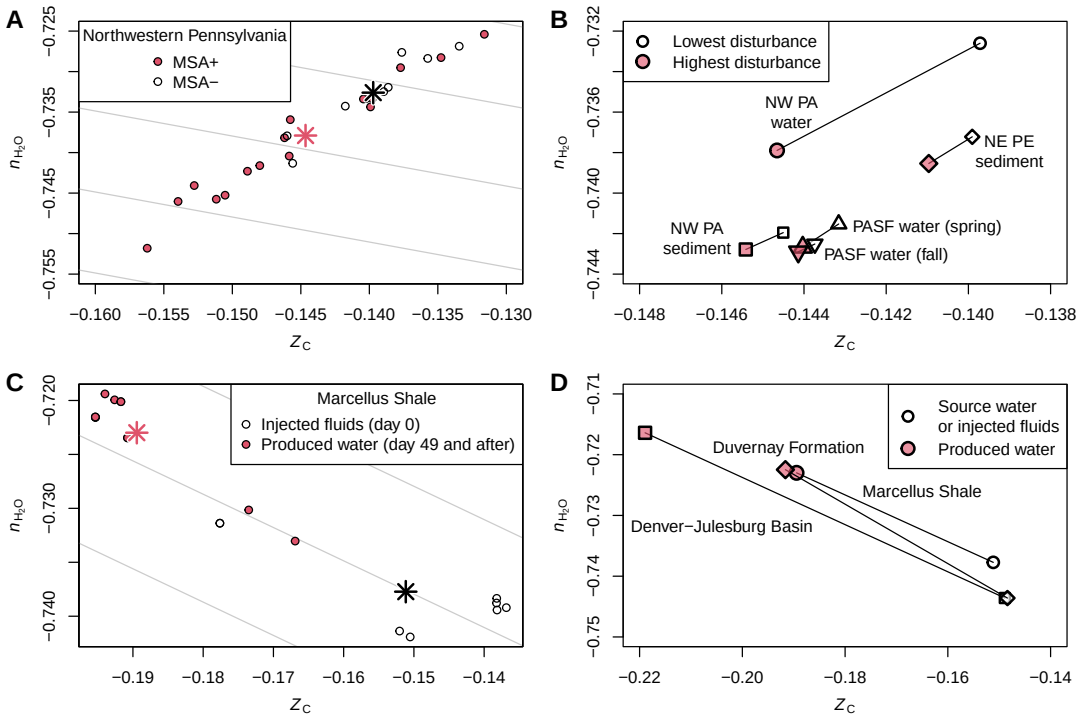

Figure 5: Decreased carbon oxidation state of inferred proteomes for communities affected by unconventional oil and gas extraction.

Community proteomes were inferred from 16S rRNA gene sequencing of UOG produced water and affected streams and sediments. (**A**) Water samples from streams affected by Marcellus Shale activity (MSA+) and non-affected streams (MSA-) in Northwestern Pennsylvania (*41*); star-shaped symbols represent group means. (**B**) Mean values for sample groups in various studies on streams in Pennsylvania: Northwestern Pennsylvania (water and sediment samples) (*41*), Northeastern Pennsylvania (sediment samples) (*42*), and Pennsylvania State Forests (PASF; water samples in spring and fall) (*35*). (**C**) Injected fluids and produced water from a hydraulically fractured well in the Marcellus Shale (*43*); star-shaped symbols represent group means. (**D**) Mean values for sample groups in various studies on produced water compared to injected fluids or source water for hydraulically fractured wells in the Marcellus Shale (*43*), Denver–Julesburg Basin (*44*), and Duvernay Formation (*38*).

Table 1: Sources of data used in this study.

Description	BioProject	Description	BioProject
Guerrero Negro mat	N/A (*25*) (**a**,**b**)	*UOG-affected streams*	
O_2 gradient (in mat)			
Yellowstone hot springs	PRJNA207095 (*26*)	Pennsylvania Streams	PRJNA394724 (*41*)
Archaea and Bacteria		Water and sediment	
ETNP water	PRJNA263621 (*31*)	Pennsylvania Streams	PRJNA449552 (*42*)
O_2 gradient		Sediment	
Baltic Sea water	PRJEB1245 (*27*) (**b**,**c**)	Pennsylvania Streams	PRJNA544240 (*35*)
Salinity gradient		Water	
Lake Fryxell mat	PRJNA291280 (*23*)		
O_2 gradient (in water)		*Source water, fracturing fluid, and produced water*	
Tibetan Plateau lakes	PRJNA294836 (*29*)		
Salinity gradient		Marcellus Shale	PRJNA229085 (*43*)
Manus Basin vents	PRJEB15554 (*24*) (**b**)		
O_2 and T gradient		Denver–Julesburg Basin	PRJNA438710 (*44*)
Qarhan Salt Lake	PRJNA388250 (*28*) (**d**)		
Saline and normal soils		Duvernay Formation	PRJNA407226 (*38*)
Ursu Lake	PRJNA395513 (*33*)		
O_2 and salinity gradient			
Black Sea water	PRJNA423140 (*22*)		
O_2 gradient			
Swiss Lakes	PRJEB27579 (*30*) (**b**)		
O_2 gradient			
Sansha Yongle Blue Hole	PRJNA503500 (*32*)		
O_2 gradient			

a. 16S rRNA gene sequences were obtained from GenBank (JN427016 to JN539989) using the `read.GenBank()` function from R package ape (*64*). For all other datasets, sequences were downloaded from the NCBI Sequence Read Archive (SRA) under the BioProject accessions listed in the table. **b**. Only bacterial 16S primers were used in these studies; other studies include both archaeal and bacterial sequences. **c**. To minimize the effects of seasonality and depth, we selected samples taken in the summer from depths of not more than 20 m. **d**. The available samples include both surface (0–10 cm) and subsurface (15–30 cm) layers. Normal (low salinity) soil samples are from outside the Qarhan Salt Lake (Water Park, Tianjin, China).

Table S1. Sequence processing statistics.

Study key	BioProject	Samples	Reads per sample (RPS) (**a**)	Filtered RPS (**b**)	Filtered read length	Single-ton % (**c**)	Chim-era % (**d**)	RPS used for classification	Classification to genus level (%) (**e**)	Map to NCBI taxonomy (%) (**f**)
Natural Environment Datasets										
HCW+13	N/A (GenBank)	10	11297	11297	(**g**)	(**g**)	18.1	9249	30	86.9
BGPF13	PRJNA207095	14	36386	17886	(**h**)	60.2	2.0	6120	20	84.4
GBL+15	PRJNA263621	15	20371	18123	250	6.1	7.0	8882	29	97.3
HLA+16	PRJEB1245	105	1878	1464	(**h**)	(**i**)	0.4	1458	37	98.0
JHM+16	PRJNA291280	8	1165067	420093 (**j**)	230	16.1	10.4	8960	17	96.3
ZLM+16	PRJNA294836	46	59934	48973	250	14.2	5.4	9456	33	97.6
MPB+17	PRJEB15554	54	165298	88931	450	54.4	7.7	8359	48	97.7
XDZ+17	PRJNA388250	29	59876	54360	250	8.9	2.4	9757	46	95.5
BCA+20	PRJNA395513	36	280205	95996	450	34.9	9.9	8638	43	98.0
SVH+19	PRJNA423140	15	6631	5718	(**h**)	11.6	5.9	4753	17	97.5
MZG+20	PRJEB27579	134 (**k**)	47294	32779	450	59.3	4.0	9151	54	94.6
HXZ+20	PRJNA503500	21 (**l**)	92727	75235	440	34.3	2.3	9773	38	95.4
Unconventional Oil and Gas Datasets										
UKD+18.sediment	PRJNA394724	93	129499	115420 (**m**)	100	10.4	1.4	9799	22	85.0
UKD+18.water	PRJNA394724	80	87546	73960 (**m**)	100	14.2	1.3	9635	21	92.7
CUN+18	PRJNA449552	29	121350	95128 (**m**)	250	23.8	4.4	9253	36	85.9
MMA+20	PRJNA544240	138	154859	114533	250	17.8	4.3	9496	38	82.4
CHM+14	PRJNA229085	46	5852	4169	(**h**)	25.4	35.2	2014	96	97.4
HRR+18	PRJNA438710	9	388424	175069 (**n**)	300	44.4	23.5	7655	99	72.7
ZLF+19	PRJNA407226	6	64644	32638	290	16.8	0.6	9938	66	97.5

a. Paired forward and reverse reads are counted as one. Paired-end reads were merged with "vsearch -fastq_mergepairs" with default options; reads that failed merging were excluded from the subsequent analysis. All "reads per sample" columns are averages of all samples.

b. Filtering was done with "vsearch -fastq_filter" with the options "-fastq_maxee_rate 0.005" and "-fastq_trunclen *length*" with length value given in next column.

c. Reads in all samples (runs) were pooled and singletons (sequences appearing exactly once) were identified with "vsearch -derep_fulllength" with the option "-maxuniquesize 1". The singletons were removed from the pooled file using "seqtk subseq"; the remaining sequences for each run were extracted from the pooled file using an awk script. After removing the singletons, 10000 sequences were subsampled from each run for the subsequent analysis using "vsearch -fastx_subsample" with the options "-sample_size 10000" and "-randseed 1234"; subsampling was not done on runs with fewer than 10000 sequences remaining after filtering and singleton removal.

d. After subsampling, runs were pooled again and chimeras were removed using "vsearch -uchime_ref" with the option "-nonchimeras" to output non-chimeric sequences (i.e. those not classified as either chimeras or borderline chimeras). The remaining sequences for each run were extracted from the output and used for taxonomic classification.

e. Classification using RDP Classifier. Any samples with < 500 classified sequences at all taxonomic levels were excluded from downstream analysis.

f. This shows the percentage of successful mappings (exact matches of rank and taxonomic name, with exceptions noted in the text) between RDP and the NCBI taxonomy for all sequences classified at phylum and lower levels.

g. FASTA sequences were downloaded from GenBank. No quality filtering or singleton removal was done.

h. For these 454 sequencing experiments, filtering was done with "vsearch -fastq_filter" with options "-fastq_stripleft 18" to remove the primer sequence and "-fastq_minlen 200 -fastq_maxlen 600 -fastq_truncqual 15" for read length and quality filtering.

i. Because of the low number of sequences, singletons were not removed.

j. After merging, reads were filtered only on length, not quality scores.

k. Library "c" was chosen, representing 134 out of 536 total runs, because it has more sequences per run than the other available libraries (a, b, and d).

l. Sequences from Station C4 were used for samples outside the Blue Hole.

m. Only forward reads were used, as indicated by the previous authors.

n. Only forward reads were used because a large majority of merge attempts failed for one or more runs (likely because of low-quality reverse reads).

Table S2. Most abundant unmapped taxonomic groups. These groups were identified by the RDP Classifier (abundances given in parentheses) but could not be mapped to the NCBI taxonomy. These groups are printed by the `getmap()` function in the JMDplots package.

Study key	Site	Most abundant unmapped groups
HCW+13	Guerrero Negro mat	genus_Potamolinea (2.27%), genus_Saccharicenans (1.54%)
BGPF13	Yellowstone hot springs	genus_Armatimonadetes_gp7 (4.05%), genus_Diapherotrites Incertae Sedis AR10 (3.91%)
GBL+15	ETNP water	phylum_Woesearchaeota (0.95%), genus_Pacearchaeota Incertae Sedis AR13 (0.47%)
HLA+16	Baltic Sea water	family_Thalassobaculaceae (0.92%), genus_Cryptomonadaceae (0.48%)
JHM+16	Lake Fryxell mat	genus_Acidibacter (0.43%), phylum_Woesearchaeota (0.4%)
ZLM+16	Tibetan Plateau lakes	genus_Cryptomonadaceae (1.1%), phylum_Woesearchaeota (0.42%)
MPB+17	Manus Basin vents	genus_Parcubacteria_genera_incertae_sedis (0.56%), genus_Thioprofundum (0.28%)
XDZ+17	Qarhan Salt Lake soils	genus_Candidatus Nanosalina (1.59%), genus_Gp4 (0.46%) (**a**)
BCA+20	Ursu Lake	genus_Desulfonatronobacter (1.53%), class_Subdivision3 (**c**) (0.24%)
SVH+19	Black Sea water	genus_Pacearchaeota Incertae Sedis AR13 (0.52%), genus_Cryptomonadaceae (0.36%)
MZG+20	Swiss Lakes	genus_Cryptomonadaceae (1.12%), genus_GpXI (1.07%) (**b**)
HXZ+20	Sansha Yongle Blue Hole	family_Arcobacteraceae (1.6%), genus_Parcubacteria_genera_incertae_sedis (0.84%)
UKD+18.sediment	Pennsylvania Streams	genus_Gp2 (5.52%) (**a**), genus_Gp3 (1.79%) (**a**)
UKD+18.water	Pennsylvania Streams	genus_Gp6 (0.91%) (**a**), genus_GpVIII (0.66%) (**b**)
CUN+18	Penn. Streams Sediment	genus_Subdivision3_genera_incertae_sedis (2.18%) (**c**), genus_Gp16 (2.1%) (**a**)
MMA+20	Penn. Streams Water	genus_Subdivision3_genera_incertae_sedis (2.96%) (**c**), genus_Gp3 (1.64%) (**a**)
CHM+14	Marcellus Shale	family_Arcobacteraceae (2.11%), order_Clostridiales (0.44%)
HRR+18	Denver-Julesburg Basin	genus_Cavicella (27.02%), genus_Gelria (0.13%)
ZLF+19	Duvernay Formation	genus_Armatimonas/Armatimonadetes_gp1 (1.14%), genus_Subdivision3_genera_incertae_sedis (0.53%) (**c**)

a. Acidobacteria.
b. Cyanobacteria.
c. Verrucomicrobia.