UC Berkeley
UC Berkeley Previously Published Works

Title

Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors

Authors

Smith, Mark B

Rocha, Andrea M

Smillie, Chris S

et al.

Peer reviewed

CrossMark
← click for updates

Natural Bacterial Communities Serve as Quantitative Geochemical Biosensors

Mark B. Smith,[a] Andrea M. Rocha,[b] Chris S. Smillie,[c] Scott W. Olesen,[d] Charles Paradis,[e] Liyou Wu,[f] James H. Campbell,[b,m] Julian L. Fortney,[g] Tonia L. Mehlhorn,[h] Kenneth A. Lowe,[h] Jennifer E. Earles,[h] Jana Phillips,[h] Steve M. Techtmann,[g] Dominique C. Joyner,[g] Dwayne A. Elias,[b] Kathryn L. Bailey,[b] Richard A. Hurt, Jr.,[b] Sarah P. Preheim,[d] Matthew C. Sanders,[d] Joy Yang,[c] Marcella A. Mueller,[h] Scott Brooks,[h] David B. Watson,[h] Ping Zhang,[f] Zhili He,[f] Eric A. Dubinsky,[i] Paul D. Adams,[i,l] Adam P. Arkin,[i,l] Matthew W. Fields,[j] Jizhong Zhou,[f] Eric J. Alm,[a,c,d] Terry C. Hazen[b,e,g,k]

Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA[a]; Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA[b]; Computational and Systems Biology Initiative, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA[c]; Biological Engineering Department, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA[d]; Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee, USA[e]; Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma, USA[f]; Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Tennessee, USA[g]; Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA[h]; Lawrence Berkeley National Laboratory, Berkeley, California, USA[i]; Department of Microbiology & Immunology, Montana State University, Bozeman, Montana, USA[j]; Department of Microbiology, University of Tennessee, Knoxville, Tennessee, USA[k]; Department of Bioengineering, University of California, Berkeley, California, USA[l]; Department of Natural Sciences, Northwest Missouri State University, Maryville, Missouri, USA[m]

M.B.S. and A.M.R. contributed equally to this article.

ABSTRACT Biological sensors can be engineered to measure a wide range of environmental conditions. Here we show that statistical analysis of DNA from natural microbial communities can be used to accurately identify environmental contaminants, including uranium and nitrate at a nuclear waste site. In addition to contamination, sequence data from the 16S rRNA gene alone can quantitatively predict a rich catalogue of 26 geochemical features collected from 93 wells with highly differing geochemistry characteristics. We extend this approach to identify sites contaminated with hydrocarbons from the Deepwater Horizon oil spill, finding that altered bacterial communities encode a memory of prior contamination, even after the contaminants themselves have been fully degraded. We show that the bacterial strains that are most useful for detecting oil and uranium are known to interact with these substrates, indicating that this statistical approach uncovers ecologically meaningful interactions consistent with previous experimental observations. Future efforts should focus on evaluating the geographical generalizability of these associations. Taken as a whole, these results indicate that ubiquitous, natural bacterial communities can be used as *in situ* environmental sensors that respond to and capture perturbations caused by human impacts. These *in situ* biosensors rely on environmental selection rather than directed engineering, and so this approach could be rapidly deployed and scaled as sequencing technology continues to become faster, simpler, and less expensive.

IMPORTANCE Here we show that DNA from natural bacterial communities can be used as a quantitative biosensor to accurately distinguish unpolluted sites from those contaminated with uranium, nitrate, or oil. These results indicate that bacterial communities can be used as environmental sensors that respond to and capture perturbations caused by human impacts.

Address correspondence to Terry C. Hazen, tchazen@utk.edu.

This article is a direct contribution from a Fellow of the American Academy of Microbiology.

With global growth in both population and affluence, the impact of human activity on the environment is widely expected to accelerate for the foreseeable future (1). Measuring the causes and consequences of these changes has become a unifying theme across many scientific disciplines, with a growing array of tools and techniques for collecting and analyzing data about the natural environment. We propose that an ideal technology should capture a wide range of useful physical and chemical properties and incorporate the results into a common format that can be quantitatively measured at low cost. Bacterial communities meet these specifications because they continuously sense and respond to their environments, forming a ubiquitous environmental surveillance network that can be inexpensively digitized through DNA sequencing. Here we seek to determine whether and how information encoded in bacterial communities can be tapped to quantitatively characterize the environment.

Many efforts have demonstrated that specific proteins (2, 3) or even whole bacterial cells (4) could be used as biosensors to trans-

late environmental signals into machine-readable data (5, 6). However, these systems must be carefully engineered before use and cannot be deployed in environments that are unsuitable for the particular proteins or cell lines being utilized. Rather than using a single macromolecule or strain, we propose integrating information gathered by native bacterial communities containing billions of cells from thousands of taxonomic groups to evaluate environmental conditions.

We propose that ecological forces predictably restrict or promote the growth of characteristic taxa in accordance with environmental conditions, a basic hypothesis that is central to ecological theory (7). Consistent with this model, previous efforts have uncovered correlations between the composition of bacterial communities and environmental features such as pH (8) or temperature (9). These and many other descriptive efforts are based on correlations fit directly to observed data rather than on the cross-validated models needed for predictive use (10), leaving indigenous biosensors largely unexplored. The recent application of machine learning to high-throughput sequence data from the human microbiome suggests that these statistical techniques might be successfully translated to assess environmental contamination (11–15).

Perturbations caused by human activity provide ideal opportunities to evaluate the predictive power of bacterial communities in response to environmental change, because human interventions cause sharp environmental gradients among sites that are otherwise very similar. As an extreme example of these human perturbations, we chose to study the Bear Creek watershed in Oak Ridge, Tennessee, a crucial site for the early development of nuclear weapons under the Manhattan Project (16). As a result of the unusual industrial processes that occurred at this site, Oak Ridge harbors spectacular geochemical gradients (17). For example, we observed in this study that pH varies by up to 7 units over less than 30 m at some locations.

RESULTS

Model and sample selection. In order to develop a robust statistical framework for integrating 16S rRNA biomarker data into a predictive model, we assessed a suite of nine machine learning tools against a benchmark classification task of distinguishing between contaminated and uncontaminated monitoring wells using data collected from this site. After integrating empirical model performance with a review of previous efforts aimed at host-associated communities (11, 18–20), we selected random forest as the best tool to identify contamination in this study (see Fig. S2 to S5 in the supplemental material for benchmark performance) (21).

To inform our statistical models of contamination, we collected extensive geochemical data and extracted DNA for sequence analysis from groundwater samples. Given the technical challenges associated with safely sampling from a nuclear waste site, we sought to maximize the geochemical diversity captured from our available sampling effort. We analyzed 25 years of monitoring data collected on 15 parameters from 812 wells across the watershed and grouped similar sites together using k-median clustering. Of the initial 100 resulting clusters, 93 were accessible for sampling (see Fig. S1 and Table S1 in the supplemental material). Groundwater from each site was accessed at a mean depth of 18.9 m (2.5 to 166.7 m) using existing monitoring wells throughout the watershed. At each of these selected sites, we measured 38 geochemical and physical features (see supplemental tables at https://sites.google.com/a/lbl.gov/enigma-extranet/pubs-review/100-well-genome-survey).

Bacterial communities within each well were collected on a 10-μm-pore-size filter to retain particle-attached cells and a 0.2-μm-pore-size filter to capture mostly free-living cells. DNA was extracted from each sample, and the 16S rRNA gene was amplified and sequenced to an average depth of 38,000 reads per sample. Despite the use of a distribution-based clustering algorithm (22) to reduce redundancy, we observed 26,943 unique operational taxonomic groups, 9,306 of which had not been previously characterized, highlighting the unusually high biological diversity of the site. Prior to prediction of contamination, we filtered low-abundance and narrowly distributed taxa, yielding 2,972 operational taxonomic units (OTUs) as features.

Accurate classification of contamination at nuclear waste site using 16S rRNA data. We found that our contamination classifier—trained on 16S rRNA data alone—is able to accurately distinguish between uncontaminated sites and those contaminated with either uranium (F1 score = 0.88) or nitrate (F1 = 0.73). Here we define contamination status based on the U.S. standards for safe drinking water for uranium and nitrate. Figure 1 shows the distribution of wells sampled across the contaminant gradients as well as classifier performance at each site. Despite nearly equal representations of features from both free-living (0.2-μm-pore-size filter, 1,554 taxa) and particle-attached (10-μm-pore-size filter, 1,418 taxa) communities, the preponderance (88%) of features important for predicting uranium was from the free-living fraction ($P < 10^{-6}$, Fisher's exact test). These results suggest that the particle-attached community may hold exclusive microniches not experienced by free-living organisms. A similar filter-specific effect is not observed for nitrate classification, suggesting that the distinction is specific to the biology of uranium-responsive taxa. As further evidence of ecological stratification across the two tested size fractions, we are able to accurately predict size fraction from otherwise identical samples drawn from the same wells using our supervised machine learning approach (see Fig. S6 in the supplemental material). The ability to both utilize and deduce ecological structure is a useful feature of this statistical approach that can be further explored as larger datasets become available.

One intrinsic limitation of this approach is that contaminated sites tend to be in close geographical proximity. As a result, it may be possible to predict contamination with just a few geographically limited strains. To control for this potential confounding effect, we retrained both classifiers, leaving out nearest geographical neighbors from the training set, and found that performance was not significantly impacted (see Fig. S7 to S9 in the supplemental material).

Instead, these models seem to discover and take advantage of genuine ecological associations. Both uranium and nitrate serve as potential electron acceptors under anoxic conditions, and many of the bacterial taxa that are most important for classification have known associations with the contaminants that they predict. For example, *Brevundimonas* species are among of the most informative features for nitrate classification and are known to be active nitrate reducers (23). Similarly, the most important features for identifying uranium contamination included species of the *Rhodanobacter* and *Rhodocyclaceae* genera, both of which have been previously identified for their role in uranium reduction and bioremediation (24). These results suggest that direct ecological
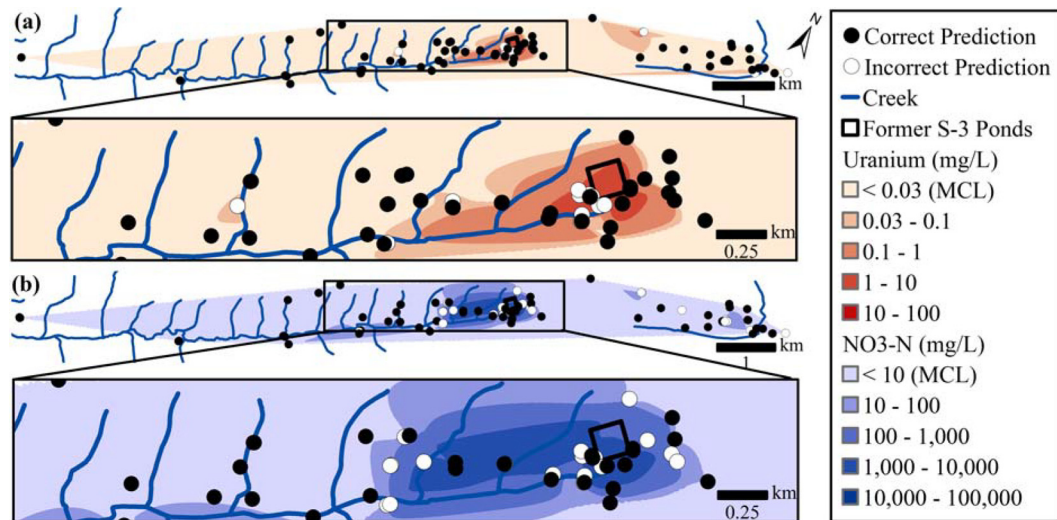
FIG 1 Uranium and nitrate contamination can be effectively identified using bacterial DNA. We trained a random forest classifier using 16S abundance data from 2,972 operational taxonomic units measured across 93 wells. Classifier performance data for uranium (a) and nitrate (b) across the Oak Ridge Field Site are shown. The maximum contaminant level (MCL) is the cutoff used to determine which sites are contaminated (samples below the cutoff are uncontaminated). Contaminant levels are measured at each well and linearly interpolated between wells. Overall classification performance values measured by specificity, sensitivity, and accuracy for detecting contamination were higher for uranium (0.71, 0.87, and 0.82, respectively) than for nitrate (0.81, 0.63, and 0.70).

associations can be used to accurately identify environmental contaminants.

Quantitative prediction of diverse geochemical targets. Many compounds that we would like to characterize with this technique may not be ecologically significant, precluding prediction through direct association. However, as a result of cross-correlations embedded in site geochemistry, it may be possible to use DNA to predict geochemical features that have only weak ecological associations (e.g., aluminum) but that are correlated with other features with strong ecological associations (e.g., pH). It seems plausible to build predictive models from these indirect associations, because many geochemical correlations are robust and emerge from physical laws. For example, the presence of dissolved oxygen (DO) directly informs oxidation-reduction (redox).

To test whether natural bacterial communities can be used as more-general geochemical biosensors, we expanded our modeling efforts beyond contamination classification to predict the values for 38 geochemical parameters measured at each site. Highlighting the flexibility of our approach, we predicted the quantitative values of each parameter at each well rather than classifying the values into discrete categories. As expected given its important role in cell physiology, we found that 16S rRNA can be used to predict pH, recovering spatial variance across the site (see Fig. 2), with a significant correlation between predicted and true values ($P < 10^{-10}$, $\tau = 0.46$, Kendall tau rank correlation). Of a total of 38 geochemical measurements, our predictions are significantly accurate for a wide range of 26 measurements ranging from manganese, a critical cofactor for many enzymes, to aluminum, which has a more limited important role in biological systems ($P < 10^{-10}$ and $P < 0.005$, respectively, Kendall tau rank correlation) (Fig. 2). Although biologically relevant traits such as pH can be predicted more accurately than traits with less-direct ecological impacts, we found that natural bacterial communities create a broadly informative imprint of their environment.

Classification of oil contamination in the Gulf of Mexico. To explore whether this approach can be applied in other ecosystems and perturbations, we analyzed previously reported data (25, 26) collected before and after the 2010 Deepwater Horizon oil spill in the Gulf of Mexico. In the worst marine oil spill in United States history, 4.1 million barrels of crude oil were released 1,500 m below the surface, 80 km from the Louisiana coast, over 85 days. Seven samples were measured in this basin before the oil spill, and 13 samples were measured at the time of the spill from locations outside the oil plume. We trained a model to distinguish between these uncontaminated samples and an additional 39 samples that were taken across a transect of the oil plume during the spill (see Fig. 3). As a demonstration of the general utility of this approach, these data were collected using an unrelated DNA measurement technology, with a PhyloChip 16S rRNA microarray, rather than by direct sequencing as described earlier. Remarkably, even with this very small training set, we are able to discriminate between contaminated and uncontaminated sites with nearly perfect accuracy (F1 score = 0.98), dramatically better than with either our uranium or nitrate classifiers (see Fig. 3).

To explore the ecological mechanisms that may underlie this surprisingly effective oil biosensor, we consider the niches of two particularly well-studied predictive features. Species of *Oceanospirillaceae* comprise a clade containing many known hydrocarbon-degrading specialists (25, 27) that are highly enriched at oil-contaminated sites. Species of *Pelagibacteraceae* comprise an oligotrophic clade that dominates nutrient-poor aquatic environments, are thought to be among the most abundant organisms on Earth, and are enriched in uncontaminated sites in our data set (28, 29). Consistent with these distinct niches, the oil biosensor is informed by both an enrichment of species of *Oceanospirillaceae* and a depletion of species of *Pelagibacteraceae* among contaminated sites. The relative abundances of organisms of these two orders alone are sufficient to accurately discriminate between contaminated and uncontaminated sites (Fig. 4B).
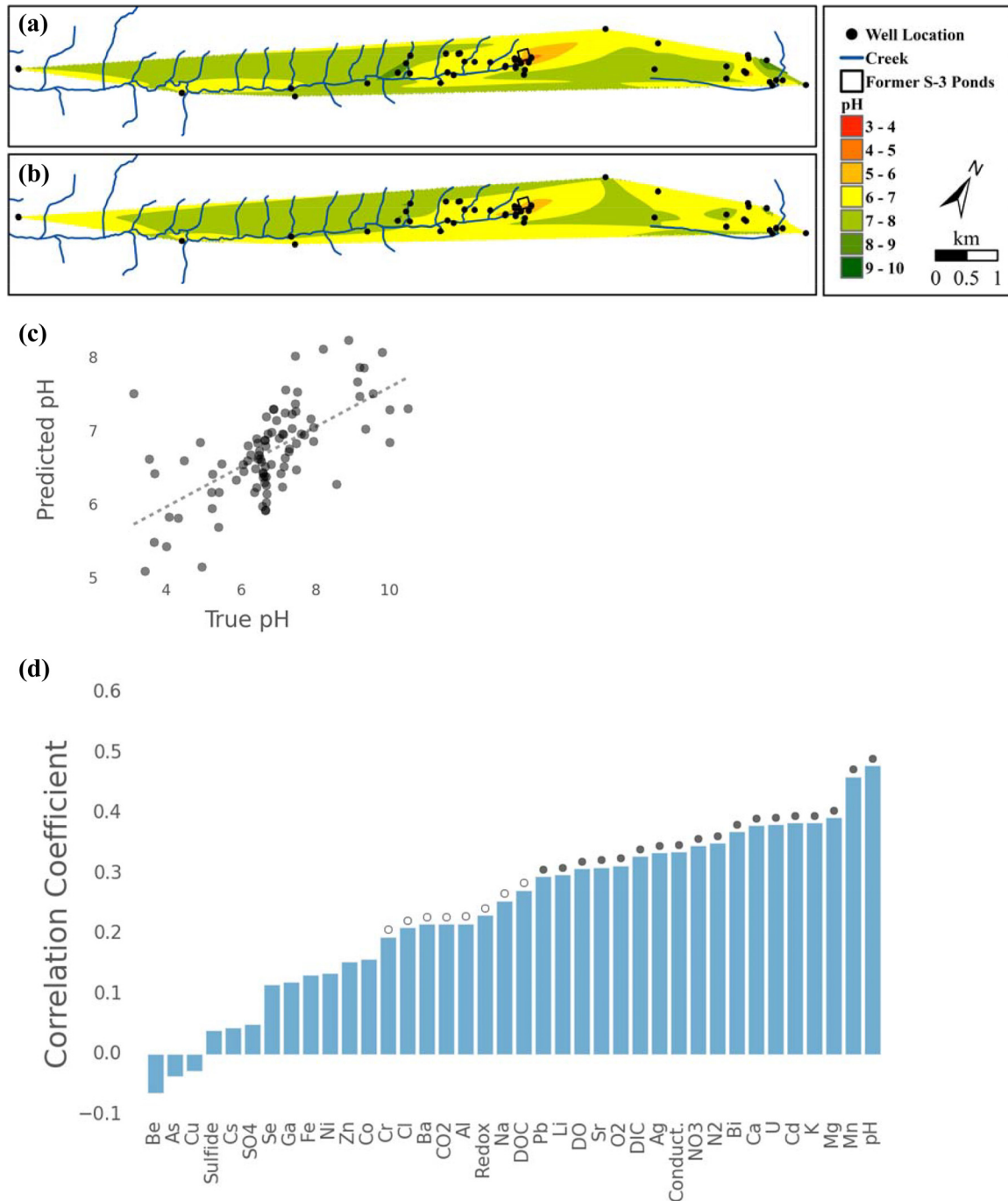
FIG 2 Bacterial DNA can be used to quantitatively predict many geochemical features. Besides classification, we can use 16S sequence data to predict quantitative values for a variety of geochemical measurements at each well. (a and b) For example, the prominent features displayed in our map of true pH (a) are recovered in our map of predicted pH (b). (C) We found that predicted values for pH are highly correlated with true values ($P < 1 \times 10^{-10}$, Kendall tau rank test). (D) We extended this approach to 38 other geochemical parameters, where we have plotted the coefficient of correlation (Kendall's tau) between true and predicted values. Among these correlations, 18 are highly significant ($P < 0.0001$, indicated by closed circles), 8 are significant ($P < 0.01$, indicated by open circles), and 12 are not significant.

Classification of previous oil contamination. Interestingly, in this plot of the oligotrophic *Pelagibacteraceae* species and the oil-degrading *Oceanospirillaceae* species, 9 samples cluster with oil-contaminated sites that were collected from within the oil plume but after hydrocarbon measurements had returned to background levels. This suggests that an ecological "memory" or imprint of previous contamination may persist, even after the contaminant has been degraded. To test this hypothesis, we used our biosensor to classify these previously contaminated sites. We are able to identify these samples with accuracy as great as that with which we
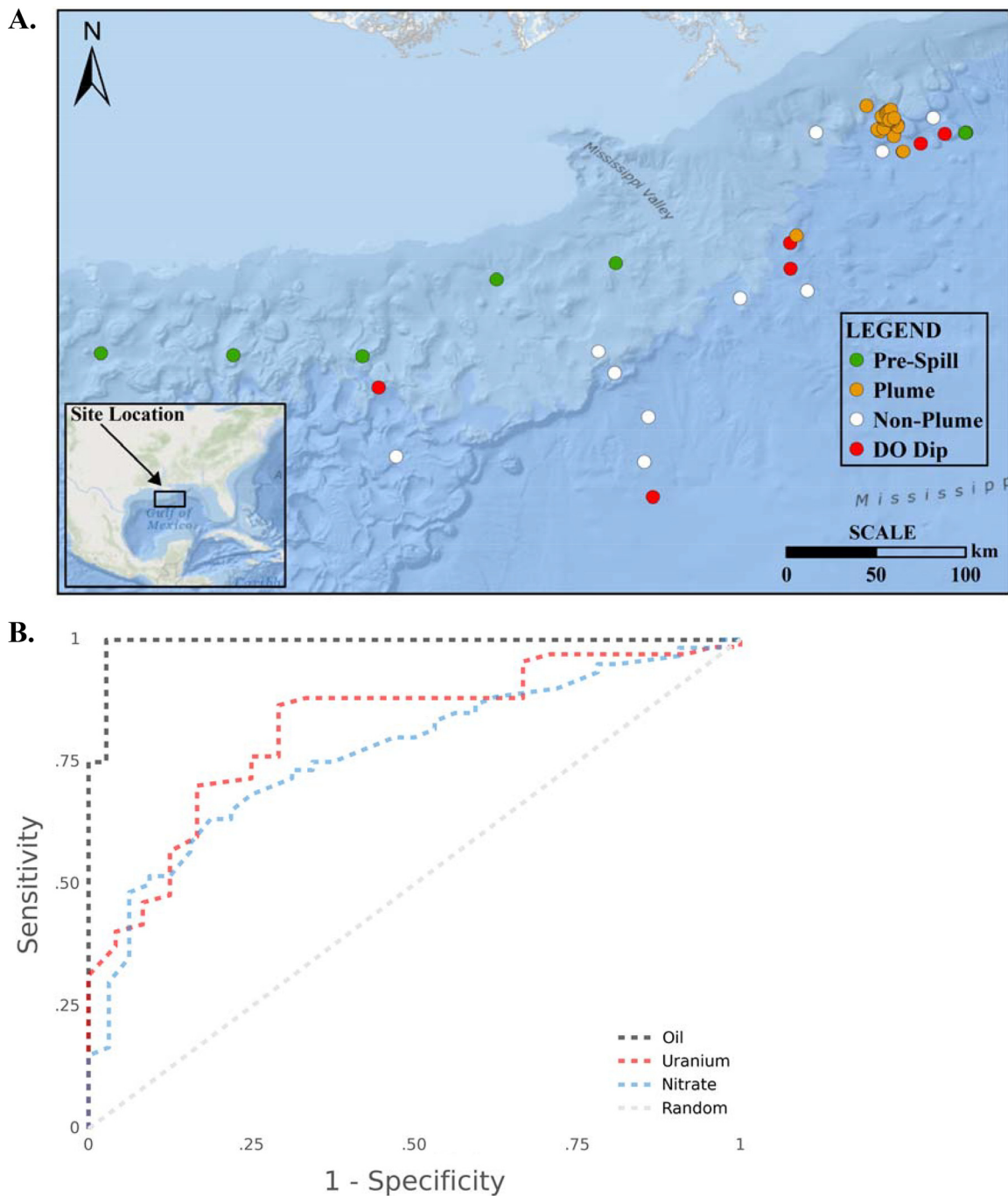
FIG 3 Near-perfect classification of oil contamination using bacterial DNA. (A) Samples collected prior to the Deepwater Horizon oil spill (green), during the spill but outside the oil plume (white), from the oil plume (orange), and from the plume but after the oil had been degraded (red) across the Gulf of Mexico are shown. (B) To compare oil classification performance with classification of uranium and nitrate, we show the receiver operator curves for all classifiers. The values for the area under the curve are 0.99 for oil, 0.82 for uranium, and 0.76 for nitrate, compared to 0.50 for an uninformative random classifier.

were able to identify truly contaminated sites (F1 =0.98) even though the oil itself is missing at these locations. These results indicate that the ecological signatures of human impacts and interventions can persist beyond the depletion of geochemical markers and demonstrate the longevity of anthropogenic perturbation in an ecological context.

To determine the phylogenetic breadth needed to build an ef-

fective indigenous biosensor, we compared the phylogenetic distributions of the most predictive features for oil and uranium, our two best classifiers (Fig. 4). There are significant phylogenetic associations between these features and the data they predict. The betaproteobacteria were enriched among uranium-predictive features ($P < 0.01$, Fisher's exact test), and the gammaproteobacteria were enriched among oil-predictive features ($P < 0.001$). How-
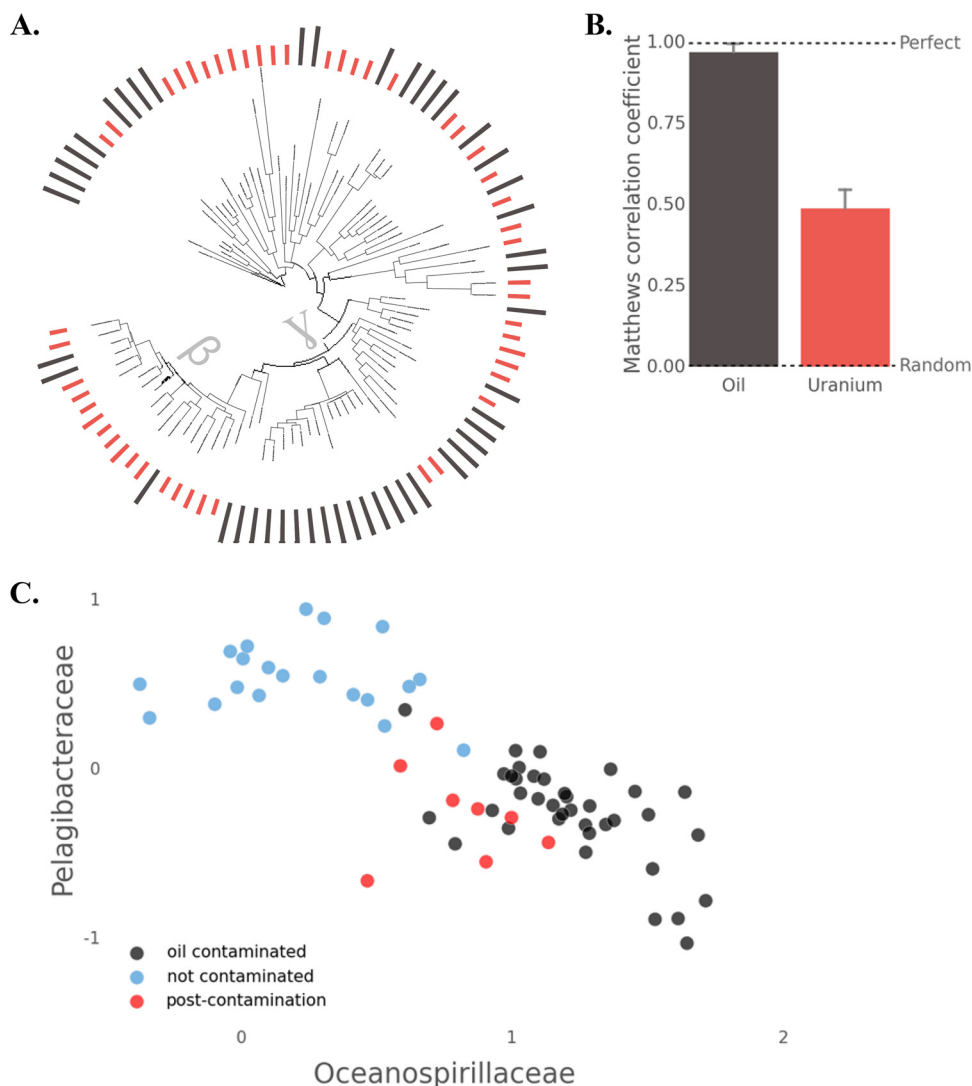
A.

B.

C.

FIG 4 random forests identify highly discriminative, biologically meaningful taxonomic groups that predict environmental conditions. (A) To understand the remarkable performance of the oil classifier, we have plotted a phylogenetic tree that includes the 50 most informative taxonomic groups for predicting uranium (red) and oil (black) data. The betaproteobacteria (β) and gammaproteobacteria (γ) clades are indicated. (B) We tested each of these features by itself as a classifier and plotted the Matthews correlation coefficient (MCC) for each of these single-feature classifiers as a bar plot at each leaf of the tree. While the best uranium features are highly informative (mean MCC = 0.49), the best features for oil classification are individually nearly perfect classifiers (mean MCC = 0.97). Error bars for the summary of these single-feature classifiers reflect 1 standard deviation. (C) The relative abundances of two highly informative features are shown for each sample. The relative abundance is expressed as the z score of each group relative to the abundances of other taxonomic groups from the same sample.

ever, beyond these groups, there is considerable variation, with features highly predictive of both oil and uranium interleaved throughout the rest of the tree of life, highlighting the phylogenetic diversity of taxa associated with these contaminants.

DISCUSSION

Previous work has explored the use of bacteria as biosensors to report information from their environments. However, these efforts have typically used electrochemical or optical properties of well-characterized strains in response to defined targets that are generally metabolized. In contrast, this study built on previous analyses of the relationships between bacterial communities and their environments (8, 30) to show that, with appropriate training data and analytical models, natural bacterial communities can be

used as biosensors for a diverse array of geochemical measurements, including many which are not directly metabolized. There is no need for prior knowledge of the relevant strains or pathways—these are identified as a product of the statistical models employed.

In this effort, we have focused on samples collected from within a single geographic area. Future efforts should prioritize the evaluation of biosensors trained in one environment against data collected from a similar environment but from a geographically distinct region. Although we found that the most informative species detected with this approach have well-defined biological associations with the geochemistry they predict, further work will be needed to explicitly evaluate whether this predictive value extends to geographically distinct sites. The ability to generalize be-

yond local biosensors will largely determine the practical utility of this approach.

Although, with existing technology, indigenous biosensors are still prohibitively costly and slow for most applications, this technical barrier seems likely to fade given the rapid pace of innovation in high-throughput molecular characterization of microbial communities. Even within the constraints of existing sequencing technology, indigenous biosensors may already be well suited to applications where it is possible to fully characterize a small training set and it is necessary to loosely monitor a broad range of geochemical features over a very large sample size.

For widespread adoption to be realized, we expect that this technique will require further reductions in the cost of sequencing to the point that the sequencing cost becomes trivial relative to the overall cost of sample acquisition and handling. In addition, this approach will require a significant expansion in the scale of curated reference datasets available. In this study, we generated our own training data, but practical utilization of this approach will require utilization of preexisting training databases to reduce the time and cost of performing this type of analysis. However, our experience demonstrates that meaningful geochemical predictions can be developed with as few as 100 reference samples.

Our observation that oil contamination can be detected even following its degradation suggests that this approach might also be favored for the detection of episodic or transient geochemical events that are difficult to capture directly, as bacterial communities may carry an embedded memory of previous exposures. The ability to detect historical geochemical exposures may become an especially valuable application, because traditional measurement techniques cannot accurately measure these historical values.

The striking results achieved with oil classification suggest that the power of a classifier is likely to scale with the strength and specificity of the selection exerted by its target. Oil is an abundant, energy-dense substrate that is unavailable for use by most organisms because of its complex chemical structure. It is a rich reward for specialists such as the members of *Oceanospirillaceae* that are able to exploit this niche. Although uranium and nitrate can serve as important electron acceptors, the ability to utilize this resource becomes ecologically relevant only in the absence of more energetically favorable electron acceptors [e.g., O_2 and Mn(III/IV)] and the presence of a suitable carbon source—rare conditions in highly oligotrophic groundwater communities. As a result, the selective advantage is less significant than that seen with oil degradation. At the same time, nitrate reduction is a general trait requiring fewer specialized genes than oil degradation. We believe that indigenous bacterial biosensors are particularly well suited to applications requiring the detection of features that, like oil, create highly specific and significant fitness effects. Previous work suggests that these principles and approaches extend beyond environmental applications and can also be employed to understand human health (14).

We expect that further development will improve this approach. These results were achieved using a single gene (16S rRNA) and relatively small training sets of fewer than 93 labeled samples. Given ubiquitous, ecologically structured gene exchange (31), we expect that many ecological associations will be captured in the flexible gene pool. Consequently, a richer set of features comprised of shotgun metagenomics, functional gene arrays, or single-cell genomes should yield more-powerful classifications. Transcriptomic data could capture instantaneous responses to en-

vironmental changes, allowing temporal tuning of the signals detected by indigenous bacterial biosensors. Larger training data sets improve model performance, making this approach more attractive with experience.

More immediately, by demonstrating the rich geochemical information captured by bacterial communities, this work supports the view that bacterial communities yield a predictable response to environmental constraints. This association between community and constraint means that bacterial systems can be used as sentinels to report the impacts of environmental perturbations and, potentially, to delineate useful stewardship strategies.

MATERIALS AND METHODS

Field data collection. (i) Site history. The Department of Energy's (DOE) Oak Ridge Field Research Center (FRC) consists of 243 acres of contaminated area and 402 acres of an uncontaminated background area used for comparison located within the Bear Creek Valley watershed in Oak Ridge, Tennessee. Contamination at this site includes radionuclides (e.g., uranium and technetium), nitrate, sulfide, and volatile organic compounds (16). The main source of contamination has been traced back to the former S-3 waste disposal ponds located within the Y-12 national security complex. During the Cold War era, these unlined ponds were the primary accumulation site for organic solvents, nitric acid, and radionuclides generated from nuclear weapon development and processing. In 1988, the S-3 ponds were closed and capped; however, contaminants from these ponds leached out, creating a groundwater contaminant plume across the field site (16). These source plumes are continuously monitored and have been the subject of a number of studies over the years (24, 32). Further information regarding the plume and sources of contamination can be found at http://www.esd.ornl.gov/orifrc/.

(ii) Well selection. We sought to maximize the impact from our limited sampling capacity by analyzing historical data collected from Oak Ridge to sample the maximum geochemical diversity of this site without exhaustively sampling all available wells.

As a response to nuclear contamination at Oak Ridge, the Department of Energy installed a constellation of monitoring wells as described above to regularly measure contamination levels across the reservation. We were able to access historical monitoring data from 834 of these monitoring wells. Regular sampling at some of these monitoring wells dates back to 1986, providing a rich time series of the site geochemistry to inform well selection. Available historical measurements from this site include copper, beta activity, alpha activity, molybdenum, sodium, potassium, uranium, sulfate, manganese, calcium, iron, nitrate, pH, chloride, and conductivity levels.

We determined that our team could sample up to 100 wells. With a target effort level in mind, we formulated well selection as a k-centroid clustering problem (with $k = 100$). Given the variance of the data, we decided to use k-median clustering to collapse the entire available well set into groups of wells that capture the geochemical diversity at the site. Figure S1 in the supplemental material illustrates the high diversity of the wells selected for study relative to all available wells. The distribution of pairwise Euclidean distances measured across the 15 available geochemical parameters for all pairs of wells is shown. Geochemical features were normalized to unitless metrics. Wells included in the study had an average pairwise distance of 1.45, while the entire population of wells had an average pairwise distance of 1.11 (arbitrary units, p <1e − 10, Mann-Whitney U test).

This clustering approach was of great practical utility given the difficulty in accessing some wells due to national security and radiation safety concerns. Because each cluster reflects wells with largely overlapping geochemical features, we selected wells within each cluster based on convenience. This enabled us to exclude especially dangerous or otherwise restricted sites from our sampling effort while preserving a systematic, principled sampling strategy. There were 7 clusters that were not sampled because all wells in the cluster were either damaged or inaccessible. The 93

clusters that were sampled were carefully selected to capture the geochemical diversity across the site.

(iii) Geochemical and physical measurements. (a) Sample collection. Groundwater samples were collected from 93 well clusters from the Oak Ridge Field Research Site between November 2012 and February 2013. Samples collected include groundwater from both contaminated and noncontaminated background wells, with each well representing a distinct geochemical transect.

All groundwater and filtered-groundwater samples were collected from the midscreen level and analyzed to determine geochemistry and to characterize the microbial community structure. Prior to collection of samples, groundwater was pumped until pH, conductivity, and oxidation-reduction (redox) values were stabilized. This was done to purge the well and the line of standing water. Approximately 2 to 20 liters of groundwater was purged from each well. For all wells, water was collected with either a peristaltic or a bladder pump using low flow in order to minimize drawdown in the well.

A total of 38 geochemical and 2 microbial parameters were measured for each well during the course of the study. Bulk water parameters, including temperature, pH, dissolved oxygen (DO), conductivity, and redox, were measured at the wellhead using an In-Situ Troll 9500 system (In-Situ Inc., CO, USA). To ensure accuracy, dissolved oxygen and pH probes were calibrated daily and the remaining probes were calibrated monthly. Sulfide and ferrous iron [Fe(II)] groundwater concentrations were determined using the U.S. EPA methylene blue method (Hach; EPA Method 8131) and the 1,10-phenanthroline method (Hach; EPA Method 8146), respectively, and analyzed with a field spectrophotometer (Hach DR 2800). All other biological and geochemical parameters were preserved, stored, and analyzed using EPA-approved and/or standard methods (41), unless otherwise indicated. A description of the sampling and analytical methods for each parameter is provided in the following sections. Lists of geochemical and microbial measurements and summary values are provided in Tables S2 to S6 in the supplemental material found at https://sites.google.com/a/lbl.gov/enigma-extranet/pubs-review/100-well-genome-survey.

(b) Dissolved gas. Preliminary dissolved gas measurements were collected using passive diffusive samplers, which measure gas concentrations in the well over a period of time.

Dissolved gases (He, H_2, N_2, O_2, CO, CO_2, CH_4, and N_2O) were measured on a SRI 8610C gas chromatograph (GC) with argon carrier gas, using a method derived from EPA RSK-175 and United States Geological Survey (USGS) Reston Chlorofluorocarbon Laboratory procedures. The GC is equipped with a thermal conductivity detector (TCD) and utilizes a 30′ Hayesep DB 100/120 column. To measure dissolved gases, 40 ml of groundwater samples was collected in precleaned volatile organic analysis (VOA) vials with no headspace and stored upside down at 4°C until analysis. To minimize diffusion of oxygen into the VOA vials through the septa, samples were analyzed within 5 days. On the day of analysis, samples were brought up to room temperature and weighed (vial plus cap plus groundwater). A 10% headspace was created by injecting argon gas via syringe into the vial while displacing an equal amount of groundwater into a second syringe. Next, the samples were shaken for 5 min and vials reweighed with the headspace. Gas samples were withdrawn using a gas-tight syringe (within 3 min after shaking has stopped). The sample was injected into a gas chromatograph for analysis, and peak areas were compared to known standards to calculate the quantity of each gas.

(c) Dissolved carbon. Dissolved organic carbon (DOC) and dissolved inorganic carbon (DIC) concentrations were determined with a Shimadzu TOC-V CSH analyzer (Tokyo, Japan) (EPA Method 415.1). Groundwater samples were collected in clean 40-ml precleaned VOA vials with no headspace. To determine DIC levels, the samples were placed on the autosampler and inorganic carbon was measured as CO_2 was released in the TOC analyzer. To determine concentrations of DOC, the samples were acidified with 2 N HCl and sparged with high-purity oxygen to remove the inorganic carbon. Samples were then injected onto the com-

bustion chamber of the carbon analyzer, and the resulting CO_2 was quantified as DOC. For each run, DIC and DOC standards were prepared based on previous knowledge of what was expected for the site. Standards ranged from 2 to 200 ppm and 0.5 to 100 ppm for DIC and DOC, respectively. Additionally, water and standards were included in the run as blanks. To minimize bacterial decomposition of some components within the groundwater sample, samples were stored at 4°C and analyzed within 1 week of collection. All reagents were prepared following EPA method protocols.

(d) Anions. Levels of anions (bromide, chloride, nitrate, phosphate, and sulfate) were determined using a Dionex 2100 system with an AS9 column and a carbonate eluent (U.S. EPA Methods 300.1 and 317.0). The Dionex system uses chromatographic separation and conductivity to measure concentrations for comparison with a standard curve. To determine anions, 20 ml of filtered groundwater (0.22-μm-pore-size filter unit) was collected in 20-ml plastic scintillation vials with no to little headspace and stored at 4°C until analysis. For analysis, the sample was loaded and 10 μl was injected into the instrument column. Calibration curves for each analyte were prepared using standard concentrations.

(e) Metals. Detection of metals (and trace elements) in the groundwater was determined on an inductively coupled plasma/mass spectrometry (ICP-MS) instrument (Elan 6100) using a method similar to EPA Method 200.7. For determination of dissolved elements, filtered groundwater samples (0.22-μm-pore-size filter unit) were collected in certified sterile VWR metal-free (<1 ppb for critical trace metals) polypropylene centrifuge tubes and stored on blue ice until transportation back to the laboratory. At the laboratory, 0.1 ml of each sample was divided into aliquots, placed in a new VWR metal-free tube, and diluted with 1% nitric acid solution to preserve the sample (pH < 2). A multielement internal standard was added directly to the diluted sample. A set of multielement calibration standards was prepared to cover the desired range of analysis. Next, samples were introduced into the system using a peristaltic pump and a PerkinElmer model AS-93 autosampler. To ensure quality control, a duplicate and matrix spike samples were included in every run (approximately 1 per 20 samples). Additionally, calibration standards were analyzed as "unknown" once every 10 samples.

To measure the availability of metals necessary for enzymes involved in denitrification (Mo, Cu, and Fe) and availability of toxic metals (e.g., U) within the groundwater across the site, 50 ml of groundwater was collected in acid-washed, autoclaved serum bottles with little to no headspace. The samples were shipped to the University of Georgia on blue ice and stored at 4°C until analysis. A Corning MP-3A distillation apparatus was used to produce the pure glass-double-distilled water (gddH$_2$O) used in all dilution and washing steps. Tubes used in ICP-MS analysis were acid washed by submersion in a 2% (vol/vol) solution of concentrated nitric acid–gddH$_2$O for 24 h and rinsed twice by submersion in pure gddH$_2$O for 24 h. Trace-metal-grade concentrated (70%) nitric acid (Fisher A509-212) was used in acidification of samples. To measure both the soluble and insoluble elements present, groundwater samples (6 ml) were placed into acid-washed 17-mm-by-20-mm Sarstedt polypropylene screw-cap conical tubes (62.554.002-PP) and centrifuged at 7,000 × g for 15 min at 4°C in a Beckman-Coulter Allegra 25R centrifuge. The supernatant was removed and placed into an acid-washed polypropylene tube and acidified with 120 μl (2% [vol/vol]) of concentrated nitric acid. A 6-ml volume of 2% (vol/vol) concentrated nitric acid–gddH$_2$O was added to the pellet. All samples were subjected to brief vortex mixing (30 s) and incubated at 37°C for 1 h in a New Brunswick Scientific G24 environmental incubator shaker with a shaking-speed setting of medium. All samples were centrifuged at 2,000 × g for 10 min in a Beckman Allegra 6R centrifuge at 25°C. Metal analysis of all samples was performed in triplicate using an Agilent 7,500ce octopole ICP-MS instrument in FullQuant mode and an internal standard with in-line addition and a multielement external standard curve as previously described (1). Samples were loaded via a Cetac ASX-520 autosampler. Control of sample introduction and data acquisition

and processing were performed using Agilent MassHunter version B.01.01.

(f) Direct cell counts. Bacterial biomass in groundwater samples was determined using the acridine orange direct count (AODC) method (25). For each well, 40 ml of groundwater samples was preserved in 4% (final concentration) formaldehyde and stored at 4°C. To prepare slides, 1 to 10 ml of groundwater was filtered through a 0.2-μm-pore-size black polycarbonate membrane (Whatman International Ltd., Piscataway, NJ). Filtered cells were then stained with 25 mg/ml of acridine orange (AO), incubated for 2 min in the dark, and filtered again to remove any unbound AO stain. The filters were rinsed with 10 ml of filter-sterilized 1× phosphate-buffered saline (PBS) (Sigma Aldrich Corp., St. Louis, MO), and the rinsed membrane was mounted on a slide for microscopy. Cells were imaged using a fluorescein isothiocyanate (FITC) filter on a Zeiss Axio Scope A1 microscope (Carl Zeiss, Inc., Germany).

Molecular biology. (i) DNA collection and extraction. DNA was collected by sequentially filtering 4 liters of groundwater through a 10.0-μm-pore-size nylon prefilter and a 0.2-μm-pore-size polyethersulfone (PES) membrane filter (144 mm in diameter) (Sterlitech Corporation). Filters were stored in 50-ml Falcon tubes and immediately stored on dry ice until transportation back to the laboratory. At the laboratory, samples were stored at −80°C until extraction using a modified Miller method (25, 33). For each sample, the filter was cut in half and each half was placed into a Lysing Marix E tube (reduced to 50% of the tubes; MP Biomedicals, Solon, OH). A 1.5-ml volume of Miller phosphate buffer and an equal volume of Miller SDS lysis buffer were added to each tube and mixed. Next, 3.0 ml of phenol-chloroform-isoamyl alcohol (25:24:1) and 3.0 ml of chloroform were added to each tube. The tubes were subjected to bead beating at medium to high speed for 5 min. The entire contents of the tube were transferred to a clean 15-ml Falcon tube and then spun at 10,000 × *g* for 10 min at 4°C. The upper phase (supernatant) was transferred to a clean 15-ml tube, and an equal volume of chloroform was added. Tubes were mixed and then spun at 10,000 × *g* for 10 min, the aqueous phase (~2 to 3 ml) was transferred to another tube, and 2 volumes of solution S3 (MoBio Power Soil, Carlsbad, CA) was added and mixed by inversion. A 650-μl volume of sample was loaded onto a spin column and filtered using a multifilter vacuum apparatus. This was continued until all the solution had been filtered. Next, 500 μl of solution S4 (MoBio Power Soil, Carlsbad, CA) were added to each filter, and the reaction mixture was then spun down at 10,000 × *g* for 30 s. The flowthrough was discarded and spun for another 30 s to ensure that all solutions had been filtered. Samples were recovered in 100 μl of solution S5 (MoBio Power Soil, Carlsbad, CA) and stored at −20°C.

(ii) Library preparation and sequencing. (a) PCR primers. A two-step PCR amplification method was used for PCR product library preparation to avoid the possible introduction of extra PCR bias by the Illumina adapter and other added components. Standard primers (515F [5′-GTG CCAGCMGCCGCGGTAA-3′] and 806R [5′-GGACTACHVGGGTWT CTAAT-3′]) targeting the V4 region of both bacterial and archaeal 16S rRNA genes without added components were used in the first step PCR.

To increase the base diversity in sequences of sample libraries within the V4 region, phasing primers were designed and used in the second step of the two-step PCR. Spacers of different lengths (0 to 7 bases) were added between the sequencing primer and the target gene primer in each of the 8 forward and reverse primer sets. To ensure that the total lengths of the amplified sequences did not vary with the primer set used, the forward and reverse primers were used in a complementary fashion such that all of the extended primer sets had exactly 7 extra bases as the spacer for sequencing phase shift. Bar codes were added to the reverse primer between the sequencing primer and the adaptor. The reverse-phasing primers contained (5′ to 3′) an Illumina adapter for reverse PCR (24 bases), unique bar codes (12 bases), the Illumina reverse read sequencing primer (35 bases), spacers (0 to 7 bases), and the target 806R reverse primer (20 bases). The forward phasing primers included (5′ to 3′) an Illumina adapter for forward PCR (25 bases), the Illumina forward read sequencing

primer (33 bases), spacers (0 to 7 bases), and the target 515F forward primer (19 bases).

(b) PCR amplification and purification. In the first-step PCR, reactions were carried out in a 50-μl reaction mixture consisting of 5 μl 10× PCR buffer II (including deoxynucleoside triphosphates [dNTPs]), 0.5 U high-fidelity AccuPrime Taq DNA polymerase (Life Technologies), a 0.4 μM concentration of both the forward and reverse target-only primers, and 10 ng or 2 μl (if total DNA amount was less than 10 ng) soil DNA. Samples were amplified in triplicate using the following program: denaturation at 94°C for 1 min and 10 cycles of 94°C for 20 s, 53°C for 25 s, and 68°C for 45 s, with a final extension at 68°C for 10 min.

The triplicate products of each sample from the first-round PCR were combined, purified with an Agencourt AMPure XP kit (Beckman Coulter, Beverly, MA), eluted in 50 μl of water, and divided into three new PCR tubes (15 μl each). The second-round PCR used a 25-μl reaction mixture (2.5 μl 10 × PCR buffer II [including dNTPs], 0.25 U of high-fidelity AccuPrime Taq DNA polymerase [Life Technologies], 0.4 μM of both forward and reverse phasing primers, and a 15-μl aliquot of the first-round purified PCR product). The amplifications were cycled 20 times following the program described above. Positive PCR products were confirmed by agarose gel electrophoresis. PCR products from triplicate reactions were combined and quantified with PicoGreen.

PCR products from samples to be sequenced in the same MiSeq run (generally 3 × 96 = 288 samples) were pooled at equal levels of molality. The pooled mixture was purified with a QIAquick gel extraction kit (Qiagen Sciences, Germantown, MD) and requantified with PicoGreen.

(c) Sequencing. Sample libraries for sequencing were prepared according to the MiSeq reagent kit preparation guide (Illumina, San Diego, CA) as described previously (34). Briefly, first, the combined sample library was diluted to 2 nM. Then, sample denaturation was performed by mixing 10 μl of the diluted library and 10 μl of 0.2 N fresh NaOH and incubated 5 min at room temperature. A 980-μl volume of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, the 20 pM library was further adjusted to reach the desired concentration for sequencing; for example, 800 μl of the 20 pM library was mixed with 200 μl of chilled Illumina HT1 buffer to make a 16 pM library to achieve about 700 paired-end reads. The 16S rRNA gene library for sequencing was mixed with about 10% (final concentration) Phix library.

A 500-cycle v1 or v2 MiSeq reagent cartridge (Illumina) was thawed for 1 h in a water bath, inverted 10 times to mix the thawed reagents, and stored at 4°C for a short time until use. Sequencing was performed for 251, 12, and 251 cycles for forward, index, and reverse reads, respectively, on MiSeq.

Computational analysis. (i) Data processing. (a) Initial filtering and processing. 16S sequence data generated from MiSeq were processed to overlap paired-end reads and to filter out poorly overlapped and poor-quality sequences. Sequences were demultiplexed using a combination of previously published programs and custom scripts. Custom scripts referenced below have been deposited for public use at https://github.com/spacocha/16S_pre-processing_scripts/. Initially, raw data were divided using a custom script (split_fastq_qiime_1.8pl) to facilitate parallel processing with SheRA (http://almlab.mit.edu/shera.html) (35). ASCII offset 33 was used in SHERA concatReads.pl, reflective of a shift in the fastq format for Illumina version 1.8 (--qualityScaling sanger). Overlapped sequences with a confidence score below 0.8 in the quality of the overlap alignment were removed (filterReads.pl). The fastq format was regenerated from the resulting fastq and quality files with the mothur (version v.1.25.0) make.fastq command (default parameters, including sanger ASCII offset 33 scaling) (36). Additionally, the corresponding index read for poorly overlapped read pairs was removed from the indexing file using a custom script (fix_index.pl). Demultiplexing and base quality filtering was done using split_libraries_fastq.py in QIIME (version 1.6.0), keeping only sequences with quality scores of 10 or more across at least 80% of the length of the total read (--min_per_read_length 0.8 —max_bad_run_

length 0 -q 10) with a Phred/ASCII offset of 33 (--phred_offset 33) (37). Finally, the primer sequences and any sequence outside the amplified region were removed using a custom script (remove_primers_staggered.pl).

(b) Creating OTUs. Operational taxonomic units (OTUs) were generated as previously described with either distribution-based clustering (DBC) or USEARCH (usearch_i86linux32 v6.0.307; http://www.drive5.com/) (22, 38). First, the sequences were truncated at 251 bp (truncate_fasta2.pl), dereplicating duplicate instances of the same sequence in the data (100% sequence clusters; fasta2unique_table4.pl) and generating a sequence-by-sample matrix (OTU2lib_count_trans1_3.pl) for any sequence with 5 or more counts in the data set. For DBC, dereplicated (100% clusters), filtered sequences were progressively clustered with UCLUST (http://www.drive5.com/) to 94% identity. DBC was run as previously described (22) from the 94% identity preclustered data, identifying significantly different distributions across samples between pairs of sequences to justify dividing the 94% cluster further. USEARCH OTUs were created at 97% identity (-cluster_fast -id 0.97), and an OTU-by-sample matrix was regenerated from the results with custom scripts (UC2list2.pl and list2mat_zeros.pl). The representative sequence for each OTU is the most abundant sequence within the OTU.

(c) Classification and removal of chimeras and nonspecific sequences. OTUs with representative sequences that are chimeras or nonspecific amplification products are removed before classification. OTU representative sequences were aligned with mothur align.seqs to the Silva bacterial alignment, which was trimmed to match the amplified region of the data. Any representative sequence which did not align to the full length of the trimmed alignment was removed (mothur; screen.seqs). Additionally, chimeric sequences were identified with uchime (http://www.drive5.com/) with default parameters and removed. Finally, sequences were classified using the Ribosomal Database Project classifier (version 2.3) (39).

(d) Identification of novel OTUs. To determine the fraction of OTUs at this site that have not been previously characterized, we used BLAST to analyze a representative sequence from each OTU against the most recent release of the GreenGenes 16S database (Version 13_5) using 97% identity gene clusters (40). For each OTU, we considered the highest-identity nearly full-length hit (>250 bp). If the best hit to the GreenGenes database represented at least 95% identity, we considered the OTU previously characterized. Of the 26,943 OTUs in our data set, 9,306 did not have a hit in the GreenGenes database with at least that level of identity. We consider these OTUs to be novel. We chose the 95% cutoff as a conservative alternative to the typical 97% cutoff used for identifying OTUs. A 97% cutoff yielded 13,371 novel OTUs. A 93% cutoff yielded 5,925 novel OTUs.

(ii) Machine learning. (a) Algorithm selection. In order to determine the most appropriate model for this application, we ran an experiment that compared eight popular machine-learning algorithms, as well as one "dummy" model. The dummy model simply reports the median value for all wells as the predicted value for each well. For our experiment, we chose to task the models with predicting the measured pH from wells that were sampled at the Oak Ridge Field Site using only 16S data from those wells. We chose to use the popular scikit-learn machine learning toolkit (18) to run the experiment, as this enabled us to quickly swap a variety of models using a common interface. As a result of our experiment, we determined that the random forest learning model fit our needs the best. Random forest had the best overall performance in terms of training time, cross-validated accuracy. We also considered that random forest has been widely used in the literature (14, 15) and has relatively few parameters to tune.

With the exception of the two linear models (Elastic Net and Lasso), each of the models we had selected performed better than the simple dummy regressor. Additionally, of all classes of learners, tree-based ensemble learners such as random forest and gradient tree boosting were the clear winners in terms of accuracy and distribution. The nonensemble decision tree method proved to be better than the linear models on aver-

age, at the expense of having a distribution that skewed toward poorer results. AdaBoost did have the single lowest error result; however, it took significantly longer to train than any other model.

(b) Data filtering. Prior to analysis, we remove OTUs that are not found in at least 20% of all samples, yielding 1,555 OTUs from the 0.2-μm-pore-size-filter data set and 1,419 OTUs from the 10-μm-pore-size-filter data set. We concatenate these matrices to allow information about the niche-specific abundance of each of these OTUs to inform our model.

(c) Random forest. After selecting Random Forests as a suitable model using the scikit-learn python module, we proceeded to use the random forest package implemented for R (random forest version 4.6-7) as our machine learning tool for all other results reported in the main text and in the supplemental material. All reported results are trained with 1,000 trees. Reported accuracies reflect the out-of-bag error for each run of the random forest model. Performance metrics are computed from a confusion matrix populated by out-of-bag predictions. Receiver operating characteristic curves are computed for classification problems using the ratio of votes for each category. Correlation coefficients for regression problems reflect the reported out-of-bag predictions relative to the true measured values. Feature importance is assessed using the native importance flag in the random forest package.

(iii) Permutation testing. To validate our machine learning pipeline, we subjected a subsample of predictions to a permutation test. Specifically, we randomized the labels associated with real training data and performed all downstream analysis as usual to determine whether our predictions could be explained by chance due to some inherent structure in the data. We performed this control to observe the variance in predictions achieved by shuffled data and analyzed real data to determine whether it would be necessary to pool results across multiple predictors to achieve reliable, replicable results.

For this test, we selected the task of classifying which wells are contaminated with uranium using our complete 16S data set (data from both the 0.2-μm-pore-size and 10-μm-pore-size filters). We chose this as our benchmark as it is a central claim of the paper and preliminary analysis indicated that these predictions were the most variable across runs. We retrained a random forest 100 times using either shuffled data or real data and computed the area under the receiver operator curve (AUC) for each replicate.

As expected, the AUC achieved with shuffled data is very close to the $x = y$ line (AUC $= 0.5$) expected by chance. The mean AUC for shuffled data is 0.49, with a standard deviation of 0.12. The AUC achieved for the real data has a higher mean (0.85) with a much lower standard deviation (0.008). These distributions are plotted below in Fig. S4 in the supplemental material.

The higher variance for shuffled data can be understood as a consequence of the random permutation of labels. In some cases, the shuffled labels happen to match the true labels, allowing a high AUC value; however, on average, the random association between labels and the training data does not allow accurate classification.

(iv) Evaluating geographic confounds. (a) Geographic structure at Oak Ridge. As illustrated in Fig. 1 in the main text, there is considerable geographic structure for the wells that were sampled at the Oak Ridge Field Site. A few significant contaminant plumes dominate the geochemical gradients measured at this site. As a result, geochemical gradients are intrinsically confounded by the geography of the site. This is a general problem for detection of contaminant dispersion from point sources. In these cases, wells that are chemically similar are likely to be geographically close.

Given this geographical confound, it is important to determine whether the biological models that we have constructed are detecting geographic or chemical signals. Although the models are not directly exposed to any geographic information, it is possible that these models could classify contaminants based on overfitting to a few taxonomic groups that are geographically restricted by chance. This interpretation would run counter to the interpretation that we present, which is that

geographic restriction is instead driven by selection from the underlying geochemistry.

(b) Data filtering. Geographic overfitting is most likely among taxa that are geographically restricted. As one methodological control against this type of overfitting, we prefilter the OTUs used as features in all of our models, excluding taxa that are not above the detection threshold in at least 20 wells. Thus, taxa used as features must be reasonably widely distributed.

(c) Evaluating the relationship between feature proximity and geographic distance. As a first step toward evaluating the effect of geography on contaminant classification, we have computed the feature-space proximity for all wells using the random forest package (21). The similarity of each pair of wells is computed based on the frequency with which these wells are found on the same terminal nodes within the forest. This is a metric of the similarity between two wells in the feature space. In Fig. S5 in the supplemental material, we compare the feature proximity of each well pair to their proximity in geographic space. As an alternative visualization of this relationship, we have binned the feature-proximity scores into 1-km groups and present the distribution of these binned data in Fig. S6.

If the models reflect general relationships between the microbiology of these sites and their geochemistry, then feature proximity should not be well correlated with geographic distance. In contrast, a geography-driven model should show a strong negative correlation between geographic distance and feature proximity—wells that are physically close should appear close in feature proximity. Consistent with a limited geographic role, the correlation between geographic distance and feature proximity is actually weakly positive for both our nitrate and uranium classifiers. Wells that are more similar in feature space are actually slightly more distant on average in geographic space. The kendall-tau correlation coefficients are 0.08 and 0.12 for nitrate and uranium, respectively.

(d) Geographic sensitivity analysis—evaluating the assumption of well independence. To directly evaluate the role of geographic proximity in our models, we performed a sensitivity analysis based on geographic exclusions and created a simple nearest-neighbors model as a null against which to compare these results. A general assumption of supervised machine learning tools such as random forest is that the training examples are independent of the test sets that are being evaluated. This assumption is violated to some degree in any environmental sampling effort, and it is difficult to determine *a priori* at what spatial scale samples might stop behaving independently.

Here we endeavored to empirically determine sample independence by evaluating performance after exclusion of geographically proximal wells. Our baseline model uses the full data set for training. We subsequently trained new random forests for each well with customized training sets that excluded wells within a defined radius of the target well. We vary the size of this radius of exclusion from 0 to 450 m. Performance decreases as this radius increases (see Fig. S7 in the supplemental material). However, it is difficult to interpret the significance of this observation without paired observation using a null model based purely on geographical proximity.

(e) A nearest-neighbor null model for evaluating geographic sensitivity. To this end, we created a simple nearest-neighbors model that predicts the label for a target well based on the labels of the k nearest other wells. We found that this model performs best when k is set to 1, so that the inferred label is set to the nearest neighbor (see Table S1 in the supplemental material). As expected given the significant geographic structure of this site, this nearest-neighbors model performs well, correctly predicting 86% and 77% of well labels for the nitrate and uranium contamination problems, respectively. However, this nearest-neighbors model is highly sensitive to the same geographic exclusion procedure applied to our random forest model described above (see Fig. S7). This suggests that although the random forest model is sensitive to geographic exclusion, the effect is much smaller than that expected from a geography-only model.

Given the low correlation between feature and geographic proximity and the modest sensitivity to geographic exclusion, we conclude that the random forest classifiers we have created are likely to reflect general biological-geochemical relationships rather than simply reflecting the geographical positions of wells within the sampling area.

(v) Geospatial analysis for data visualization. Geospatial analysis was performed using ArcMap 10.1 software by Environmental Systems Research Institute (esri) and displayed using the World Geodetic System 1984 (WGS 1984) coordinate system. The latitude and longitude of groundwater well and marine station locations were uploaded to ArcMap along with measured and predicted analyte concentration data to create point shapefiles. The point shapefiles of measured or predicted analyte concentrations in groundwater were interpolated using the Natural Neighbor technique within the Spatial Analyst Tools of the ArcToolbox. Point concentration data were used as the input point feature for the z value field. The remaining input parameters were set to default settings. The output of the interpolation resulted in floating point raster files consisting of 470 columns by 250 rows with a square pixel size of 2.1E-4 by 2.1E-4 degrees. The line shapefile for the surface water bodies at the Oak Ridge Reserve (ORR) was provided by the United States Geological Survey (USGS) National Hydrography data set (NHD). The basemap for the Gulf of Mexico (GOM) was designed and developed by esri with contributions from General Bathymetric Chart of Oceans (GEBCO), National Oceanic and Atmospheric Administration (NOAA), National Geographic, and DeLorme.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00326-15/-/DCSupplemental.

Figure S1, PNG file, 0.1 MB.
Figure S2, PNG file, 0.04 MB.
Figure S3, PNG file, 0.1 MB.
Figure S4, PNG file, 0.03 MB.
Figure S5, PNG file, 0.05 MB.
Figure S6, PNG file, 0.1 MB.
Figure S7, PNG file, 0.6 MB.
Figure S8, PNG file, 0.1 MB.
Figure S9, PNG file, 0.1 MB.
Table S1, DOCX file, 0.03 MB.

ACKNOWLEDGMENTS

REFERENCES

1. **Moss RH, Edmonds JA, Hibbard KA, Manning MR, Rose SK, van Vuuren DP, Carter TR, Emori S, Kainuma M, Kram T, Meehl GA, Mitchell JF, Nakicenovic N, Riahi K, Smith SJ, Stouffer RJ, Thomson AM, Weyant JP, Wilbanks TJ.** 2010. The next generation of scenarios for climate change research and assessment. Nature **463:**747–756. http://dx.doi.org/10.1038/nature08823.

2. Fischer T, Agarwal A, Hess H. 2009. A smart dust biosensor powered by kinesin motors. Nat Nanotechnol 4:162–166. http://dx.doi.org/10.1038/nnano.2008.393.

3. Wu J, Park JP, Dooley K, Cropek DM, West AC, Banta S. 2011. Rapid development of new protein biosensors utilizing peptides obtained via phage display. PLoS One 6:e24948. http://dx.doi.org/10.1371/journal.pone.0024948.

4. Belkin S. 2003. Microbial whole-cell sensing systems of environmental pollutants. Curr Opin Microbiol 6:206–212. http://dx.doi.org/10.1016/S1369-5274(03)00059-6.

5. Su L, Jia W, Hou C, Lei Y. 2011. Microbial biosensors: a review. Biosens Bioelectron 26:1788–1799. http://dx.doi.org/10.1016/j.bios.2010.09.005.

6. D'Souza SF. 2001. Microbial biosensors. Biosens Bioelectron 16:337–353. http://dx.doi.org/10.1016/S0956-5663(01)00125-7.

7. Darwin C. 1859. On the origin of the species by natural selection. Murray, London, England.

8. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111–5120. http://dx.doi.org/10.1128/AEM.00335-09.

9. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB. 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A 106:1374–1379. http://dx.doi.org/10.1073/pnas.0808022106.

10. Hawkins DM, Basak SC, Mills D. 2003. Assessing model fit by cross-validation. J Chem Inf Comput Sci 43:579–586. http://dx.doi.org/10.1021/ci025626i.

11. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. 2011. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe 10:292–296. http://dx.doi.org/10.1016/j.chom.2011.09.003.

12. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343–359. http://dx.doi.org/10.1111/j.1574-6976.2010.00251.x.

13. Beck D, Foster JA. 2014. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One 9:e87830. http://dx.doi.org/10.1371/journal.pone.0087830.

14. Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, Giannoukos G, Ciulla D, Tabbaa D, Ingram J, Schauer DB, Ward DV, Korzenik JR, Xavier RJ, Bousvaros A, Alm EJ. 2012. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLoS One 7:e39242. http://dx.doi.org/10.1371/journal.pone.0039242.

15. Metcalf JL, Wegener Parfrey L, Gonzalez A, Lauber CL, Knights D, Ackermann G, Humphrey GC, Gebert MJ, Van Treuren W, Berg-Lyons D, Keepers K, Guo Y, Bullard J, Fierer N, Carter DO, Knight R. 2013. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. eLife 2:e01104. http://dx.doi.org/10.7554/eLife.01104.

16. Watson DB, Kostka JE, Fields MW, Jardine PM. 2004. The Oak Ridge Field Research Center conceptual model. https://public.ornl.gov/orifc/FRC-conceptual-model.pdf.

17. Fields MW, Bagwell CE, Carroll SL, Yan T, Liu X, Watson DB, Jardine PM, Criddle CS, Hazen TC, Zhou J. 2006. Phylogenetic and functional biomarkers as indicators of bacterial community responses to mixed-waste contamination. Environ Sci Technol 40:2601–2607. http://dx.doi.org/10.1021/es051748q.

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830.

19. Yang C, Mills D, Mathee K, Wang Y, Jayachandran K, Sikaroodi M, Gillevet P, Entry J, Narasimhan G. 2006. An eco-informatics tool for microbial community studies: supervised classification of amplicon length heterogeneity. ALH profiles of 16S rRNA. J Microbiol Methods 65:49–62. http://dx.doi.org/10.1016/j.mimet.2005.06.012.

20. Tanaseichuk O, Borneman J, Jiang T. 2014. Phylogeny-based classification of microbial communities. Bioinformatics 30:449–456. http://dx.doi.org/10.1093/bioinformatics/btt700.

21. Liaw A, Wiener M. 2002. Classification and regression by randomForest. R News 2:18–22.

22. Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, Alm EJ. 2013. Distribution-based clustering: using ecology to refine the operational taxonomic unit. Appl Environ Microbiol 79:6593–6603. http://dx.doi.org/10.1128/AEM.00342-13.

23. Kavitha S, Selvakumar R, Sathishkumar M, Swaminathan K, Lakshmanaperumalsamy P, Singh A, Jain SK. 2009. Nitrate removal using Brevundimonas diminuta MTCC 8486 from ground water. Water Sci Technol 60:517–524. http://dx.doi.org/10.2166/wst.2009.378.

24. Green SJ, Prakash O, Jasrotia P, Overholt WA, Cardenas E, Hubbard D, Tiedje JM, Watson DB, Schadt CW, Brooks SC, Kostka JE. 2012. Denitrifying bacteria from the genus Rhodanobacter dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. Appl Environ Microbiol 78:1039–1047. http://dx.doi.org/10.1128/AEM.06435-11.

25. Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D'haeseleer P, Holman H-YN, Osman S, Lu Z, Nostrand JDV, Deng Y, Zhou J, Mason OU. 2010. Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science 330:204–208. http://dx.doi.org/10.1126/science.1195979.

26. Dubinsky EA, Conrad ME, Chakraborty R, Bill M, Borglin SE, Hollibaugh JT, Mason OU, Piceno M, Y, Reid FC, Stringfellow WT, Tom LM, Hazen TC, Andersen GL. 2013. Succession of hydrocarbon-degrading bacteria in the aftermath of the deepwater Horizon oil spill in the Gulf of Mexico. Environ Sci Technol 47:10860–10867. http://dx.doi.org/10.1021/es401676y.

27. Teramoto M, Ohuchi M, Hatmanti A, Darmayati Y, Widyastuti Y, Harayama S, Fukunaga Y. 2011. Oleibacter marinus gen. nov., sp. nov., a bacterium that degrades petroleum aliphatic hydrocarbons in a tropical marine environment. Int J Syst Evol Microbiol 61:375–380. http://dx.doi.org/10.1099/ijs.0.018671-0.

28. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806–810. http://dx.doi.org/10.1038/nature01240.

29. Carini P, Steindler L, Beszteri S, Giovannoni SJ. 2013. Nutrient requirements for growth of the extreme oligotroph "Candidatus Pelagibacter ubique" HTCC1062 on a defined medium. ISME J 7:592–602. http://dx.doi.org/10.1038/ismej.2012.122.

30. Schmidtova J, Baldwin SA. 2011. Correlation of bacterial communities supported by different organic materials with sulfate reduction in metal-rich landfill leachate. Water Res 45:1115–1128. http://dx.doi.org/10.1016/j.watres.2010.10.038.

31. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244. http://dx.doi.org/10.1038/nature10571.

32. Green SJ, Prakash O, Gihring TM, Akob DM, Jasrotia P, Jardine PM, Watson DB, Brown SD, Palumbo AV, Kostka JE. 2010. Denitrifying bacteria isolated from terrestrial subsurface sediments exposed to mixed-waste contamination. Appl Environ Microbiol 76:3244–3254. http://dx.doi.org/10.1128/AEM.03069-09.

33. Miller DN, Bryant JE, Madsen EL, Ghiorse WC. 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Appl Environ Microbiol 65:4715–4724.

34. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621–1624. http://dx.doi.org/10.1038/ismej.2012.8.

35. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW. 2010. Unlocking short read sequencing for metagenomics. PLoS One 5:e11840. http://dx.doi.org/10.1371/journal.pone.0011840.

36. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541. http://dx.doi.org/10.1128/AEM.01541-09.

37. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D,

Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7:**335–336. http://dx.doi.org/10.1038/nmeth.f.303.

38. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. BioInformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.

39. **Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM.** 2005. The Ribosomal Database project (RDP-II): se-quences and tools for high-throughput rRNA analysis. Nucleic Acids Res **33:**D294–D296. http://dx.doi.org/10.1093/nar/gki038.

40. **DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72:**5069–5072. http://dx.doi.org/10.1128/AEM.03006-05.

41. **American Public Health Association.** 2012. Standard methods for the examination of water and wastewater. https://www.standardmethods.org.