Ecological insights from bacterial networks

Mark Burnham Smith

A.B. Ecology and Evolutionary Biology

Princeton University, 2009

SUBMITTED TO THE MICROBIOLOGY GRADUATE PROGRAM IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN MICROBIOLOGY AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2014

Signature redacted

Signature of Author:.....

.....

Mark Burnham Smith

Microbiology Graduate Program

Signature redacted

Certified by:.....

.....

Eric J. Alm

Associate Professor of Biological Engineering

Thesis Supervisor

Signature redacted

Accepted by:.....

.....

Michael T. Laub

Associate Professor of Biology

Chair, Microbiology Graduate Program

Ecological insights from bacterial networks

Mark Burnham Smith

Submitted to the Microbiology Graduate Program on August 28, 2014
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Microbiology

ABSTRACT
Microbes occupy a wide range of important niches ranging from global biogeochemical cycles to metabolism in the human gut. Yet microbes rarely act in isolation. Instead, they thrive in complex communities with myriad combinatorial interactions. In this work I explore the nature of these bacterial networks, using computational tools to uncover ecological associations with relevance to both human health and environmental restoration.

I begin with the discovery of a massive, global network of recent gene exchange linking even distantly related bacteria from the far corners of earth. To uncover this network, I developed and validated a simple evolutionary rate heuristic and applied it to report recent transfers across nearly 5 million pairwise interactions among bacterial genomes. I interrogated this network for associations between rates of horizontal gene transfer (HGT) and differences in the geography, ecology and phylogenetic history of each pair of genomes. Of these influences, ecological overlap is the most important force shaping recent gene exchange.

In the second chapter, I use CRISPR arrays as a record of recent infections to investigate the host range of mobile genetic elements. I report 7,009 pairs of genomes that contain identical spacers and are at least 10% divergent at the 16S rRNA gene, implying an overlap in genetic element host range. This provides a mechanistic framework to understand the transfers uncovered in the first chapter.

In the final section of this work, I exploit this powerful link between bacterial communities and their environments to create a machine-learning algorithm that translates DNA from natural bacterial communities into accurate, quantitative readouts of environmental conditions. I develop this approach using 16S rRNA sequence data from 93 groundwater wells in Oak Ridge, Tennessee to predict a diverse array of 26 geochemical measurements. I validate this technique using microarray data from the Deepwater Horizon oil spill. The predictive power of these models generally emerges from the composite of the entire community and its interactions, rather than from a single strain. As a whole, this body of work demonstrates the profound connections that link the microbial world into an ecologically structured network.

Thesis Supervisor: Eric J. Alm
Title: Associate Professor of Biological Engineering

Table of Contents

Title Page 1

Abstract 2

Table of Contents 3

Acknowledgements 4

Personal Context 6

Introduction 7

Chapter 1: Ecology drives a global network of gene exchange connecting the human microbiome 11
- 1.1 Abstract 11
- 1.2 Main Text 12
- 1.3 Methods 18
- 1.4 Figures 26
- 1.5 Supplemental Figures and Tables 30

Chapter 2: Identical CRISPR spacers among distantly related bacteria reveal common strategies to target promiscuous mobile elements. 37
- 2.1 Abstract 37
- 2.2 Main Text 38
- 2.3 Methods 43
- 2.4 Figures 44

Chapter 3: Natural bacterial communities as quantitative biosensors 48
- 3.1 Abstract 48
- 3.2 Main Text 48
- 3.3 Methods 57
- 3.4 Figures 73
- 3.5 Supplemental Figures and Tables 77

Conclusions and future directions 87

Bibliography 93

Acknowledgements

I would like to gratefully acknowledge the invaluable advice and support of Eric Alm, my thesis supervisor and mentor. Eric has taught me how to identify interesting scientific problems, solve them and effectively share and apply the results. He has also become a great friend, personal trainer and adventure racing teammate. I would also like to recognize the mentorship and guidance of my thesis advisory committee members, Martin Polz, Ed Delong and Terry Hazen.

I am deeply grateful to the National Science Foundation for supporting my studies through a Graduate Research Fellowship. I am also thankful to the Martin Family Foundation for supporting my work through the Martin Fellowship and to BP for supporting me as an MITEI Energy Fellow. In addition to these fellowships, I would like to specially acknowledge the support provided by the Department of Energy ENIGMA program and the entire ENIGMA team for enabling the work presented in Chapters 2 and 3. I am also deeply grateful to the MIT Microbiology program (and it's wonderful students, faculty and staff) for supporting me throughout my time at MIT.

In addition to this formal support, I am deeply appreciative of the many contributions made by the Alm Lab and Parsons's community both directly to this work as described below and more generally to my development as a scientist and a human. Among this community, I am particularly grateful to Chris Smillie, who made numerous important contributions both to the works presented here and to many other projects beyond this dissertation. Specifically, as the co-author of the manuscript presented in Chapter 1, Chris and I worked together to develop and implement the evolutionary rate heuristic that we employed to discover and analyze recent horizontal gene transfer events. Chris also taught me many of the basic principles of machine learning and developed our lab's code-base for working with Random forest, which I extended to classify environmental contaminants in Chapter 3. Chris has been a great collaborator and friend, always happy to meet up for a late-night brainstorm, embark on an ill-conceived outdoor adventure or drink tea. I would also like to recognize the tireless work of Andrea Rocha and the

University of Tennessee and Oak Ridge National Laboratory field teams that collected all of the data that I have analyzed in Chapter 3.

Perhaps the most important contribution to this work comes from my family and particularly, my parents who always found time to engage with my curiosity growing up. I would also like to thank Carolyn for tolerating unreasonably frequent fecal-based conversations and supporting my passions, even when they lead in unorthodox and surprising directions. Finally, I would like to dedicate this work to the memory of Margaret Burnham, an educator, researcher, inspiration and my grandmother.

Personal Context

While I was interviewing for graduate school in 2009, I attended a lecture at Berkeley in which the super-exponential decline in the cost of genome sequencing was first presented to me. The presenter suggested that increasingly accessible molecular tools could be used to re-evaluate many commonly held dogmas in microbiology.

At the time, my experience as a field ecologist at Princeton had taken me from the boreal and mixed forests of Northern Wisconsin to the tropical dry forests of Panama to investigate the relationship between the structure and diversity of forest communities and their underlying environmental constraints, a theme that I will revisit in this work. I found myself fascinated by the questions of ecology and evolutionary biology but challenged by the laborious methods required to generate even poorly constrained models. During my junior year at Princeton, I shifted my focus to experimental microbiology, where I enjoyed the rapid progress possible through work with microbial systems. This experience drew me towards graduate studies in Microbiology.

At MIT, I rotated through traditional experimental microbiology labs before arriving in the Alm lab to finally explore the vague promise of next-generation sequencing, which was still entirely new to me. I quickly realized that genomics offered even richer data than experimental work and immersed myself in the computational tools needed to take advantage of this new data source. Each transition from the field to the lab to computation was motivated by my interest in finding richer data to interrogate the basic principles of ecology and evolution. Although each of these transitions required significant investments in new skills, I gained a deep appreciation for the value and basic methodologies of each domain during this journey. This experience has enabled me to effectively collaborate across traditional methodological lines as demonstrated by this work.

Introduction

Much of the early work in genomics focused on the analysis of single genomes (Himmelreich et al. 1996; Blattner et al. 1997; McClelland et al. 2001) with later efforts using comparative approaches to interrogate gene function or evolutionary history (Makarova et al. 1999; Arigoni et al. 1998; Koonin, Aravind, and Kondrashov 2000). However, given the emerging view of bacterial communities as complex interacting networks (Miller and Bassler 2001; Dubey and Ben-Yehuda 2011), this work seeks to develop genomic approaches to search for evidence of these interactions and the rules that govern them.

Horizontal gene transfer is a particularly attractive interaction to study, both because of its importance in bacterial evolution and because it leaves a detectable signature in affected genomes. HGT enables bacteria to rapidly adapt to changing environments, tapping a broad pool of potentially useful elements. Evidence from resequencing of epidemic strains suggests that HGT enables much more rapid and widespread evolutionary plasticity than mutations (Garg et al. 2003). Examining the process and product of gene acquisition provides insight into the nature of bacterial evolution. HGT is also a favorable process to study because it leaves clear imprints on the genomes involved, making it practically tractable for investigation.

Prior to this work, HGT has been widely studied as a driver of bacterial evolution. However, this earlier work primarily focused on HGT as a historical event rather than an ongoing process (Ochman, Lawrence, and Groisman 2000; Thomas and Nielsen 2005; Gogarten and Townsend 2005; David and Alm 2011). In the first chapter of this thesis I present results from the analysis of recent horizontal gene transfer events. Unlike historical transfers, the environmental and in some cases, geographical associations of the genomes involved in these cases have been preserved. As a result, with this approach it is possible to evaluate the relative impact of phylogenetic history, ecological similarity and geographical proximity on the frequency of gene transfers.

To facilitate these comparisons, I examine nearly 5 million pairs of genome interactions and find over 10,000 unique genes that are transferred. Across this network, I find that ecology is the dominant force shaping recent gene transfers, with 25-fold more HGT among human-associated isolates than among diverse non-human strains. More narrowly defined niches such as shared sub body-site, oxygen tolerance or pathogenicity are further enriched in transfer. I suggest that although there are likely more opportunities for HGT among microbes residing in similar environments, because the strains considered in this analysis are drawn from around the world, the dominant effect is likely caused by positive selection. Bacteria occupying similar niches are most likely to share overlapping selective pressures, causing the proliferation of mutually useful genes that happen to be transferred. Functional analysis of genes transferred among distantly related bacteria occupying the same niches (like association with meningitis) supports this view as many of these transferred genes are known to play important roles in the niches they are associated with (such as virulence). By developing, validating and applying a new evolutionary rate heuristic for identifying recent HGT, I uncovered a massive network of gene transfer that provides an imprint of recent ecological interactions.

In chapter one, I reveal a massive network of recent gene exchange among distantly related bacteria. This surprising observation suggests that the prevailing dogma about the narrow host range of most mobile genetic elements (MGE) could understate the promiscuity of phage and plasmids. To further investigate this question, in the second chapter, I use CRISPR arrays to probe the host range of MGE.

The inherent difficulty and limitations of culture-based assays have been one of the great challenges constraining traditional efforts to probe the host range of mobile genetic elements. With 10.1 million pairwise comparisons available from sequenced genomes that contain CRISPR, next-generation sequencing has opened up a new opportunity for systematic analysis of MGE host-ranges using comparative genomics. In this chapter, I perform a systematic search for evidence of broad-host range elements in CRISPR arrays. I find 7,009 examples of distantly related genomes (>10% 16S rRNA divergence) that share at least one identical CRISPR spacer. More than half of genome pairs that contain

identical spacers also contain identical repeats, suggesting that HGT of entire CRISPR arrays is the primary explanation for shared spacers. Spacers shared across genomes are four and three times more likely than unique spacers are to be found in phage and plasmid databases respectively. This suggests that shared CRISPR spacers may reflect selection to target especially common components of MGE. Although the genes targeted by CRISPR may travel independently from the MGE being targeted, the observation of identical spacers provides further evidence to support a broad host range for MGE, consistent with the observation of rampant HGT presented in the first chapter.

In the first chapter I use environmental associations to explain bacterial interactions. In the final chapter, I explore the inverse relationship, using bacterial interactions to make inferences about the environment. I apply machine-learning tools to predict a variety of environmental features using 16S rRNA sequence data to reconstruct bacterial networks. By training statistical models on samples across contaminated field sites, I am able to create robust predictors of environmental contaminants at these sites. I present this approach as an indigenous biosensor that can use bacteria as a ubiquitous environmental monitoring system to detect changes in the environment.

I develop this approach using field and sequence data collected at a nuclear weapons site in Oak Ridge, Tennessee that covers extreme geochemical gradients. As part of a broad collaborative effort, I directed the site selection to maximize the diversity of sampled wells in order to best inform the downstream statistical modeling of these systems. I developed models that are able to both distinguish between sites that are contaminated with the two most common pollutants at the site (uranium and nitrate) and that can quantitatively predict the value for a range of 26 geochemical measurements collected at this site.

To determine how general this approach might be, I extend this analysis to data collected from the Deepwater Horizon oil spill in 2010. In this case I use 16S rRNA data to predict which sites are contaminated by oil and which are not. To demonstrate the portability of this approach, I use a 16S rRNA microarray as an orthogonal DNA measurement

technology. I find that I am able to create a nearly perfect classifier for oil contamination. Interestingly, when I extend this approach to sites that were sampled after the oil was degraded, these sites are accurately classified as oil contaminated, suggesting that bacterial communities contain a measurable signature of previous environmental exposures. Although in most cases, the predictive power of this approach emerges from a composite view of the entire community and its interactions, in this extreme case of oil degradation, I show that even the abundance of single strains are sufficient to distinguish between contaminated and uncontaminated sites. This is likely due to the powerful selection exerted by the influx of a rich carbon source into an otherwise oligotrophic environment. I conclude this chapter by discussing the potential future applications of this approach, which uses bacterial networks to report environmental conditions.

As a whole, this work demonstrates the myriad connections that link bacterial networks into ecologically informative systems. I have used a wide array of computational tools to analyze molecular data captured from a wide range of environments around the world in this work. However, these disparate methods and sites are unified by the underlying theme of exploring connections among bacteria and between bacteria and their environments. This work sheds new light on the depth of these interactions (chapter three) and the evolutionary mechanisms that tie bacterial networks together (chapters one and two). With the increasingly widespread availability of rich genomic data, I expect that future work will continue to use the principles of comparative genomic and meta-genomic analyses presented here to further probe the nature of bacterial networks.

Chapter 1: Ecology drives a global network of gene exchange connecting the human microbiome

Smillie CS*, Smith MB*, Friedman J, Cordero OX, David LJ, Alm EJ (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**: 241-244

1.1 Abstract

Horizontal gene transfer (HGT), the acquisition of genetic material from non-parental lineages, is known to play an important role in bacterial evolution (Ochman, Lawrence, and Groisman 2000; Koonin, Makarova, and Aravind 2001). Notably, HGT provides rapid access to genetic innovations, allowing traits like virulence (Chen and Novick 2009), antibiotic resistance (Lester et al. 2006), and xenobiotic metabolism (Hehemann et al. 2010) to spread through the human microbiome. Recent anecdotal studies that provide snapshots of active gene flow on the human body highlight the need to determine the frequency of such recent transfers and the forces that govern these events (Lester et al. 2006; Hehemann et al. 2010). Through the analysis of 2,235 full bacterial genomes, here we report the discovery and characterization of a vast, human-associated network of gene exchange, large enough to directly compare the principal forces shaping HGT for the first time. We show that this network of 10,770 unique, recently transferred (> 99% nucleotide identity) genes is principally shaped by ecology rather than geography or phylogeny, with most gene exchange occurring among isolates from ecologically similar, but geographically separated environments. For example, we observe 25-fold more HGT among human-associated bacteria than among ecologically diverse non-human isolates ($P = 3.0 \times 10^{-270}$). Within the human microbiome, we show this ecological architecture continues across multiple spatial scales, functional classes, and ecological niches with transfer further enriched among bacteria that inhabit the same body site, exhibit the same oxygen tolerance, or have the same ability to cause disease. This structure offers a window into the molecular traits that define ecological niches, insight we use to uncover sources of antibiotic resistance and to identify genes associated with the pathology of meningitis and other diseases.

---

*These authors contributed equally to this work.

1.2 Main Text

The human body is a complex biological network comprised of ten microbes for each human cell and one hundred microbial genes for each unique human gene (Gill 2006). Because this hidden microbial majority is known to have profound impacts on many aspects of human health including immunity (Round and Mazmanian 2009), inflammatory disease (Xavier and Podolsky 2007), and obesity (Ley et al. 2006), considerable efforts are underway to document the genetic diversity of the human microbiome. It is unclear what role HGT plays in the generation and distribution of this biochemical repertoire, although anecdotal findings suggest that it may be significant (Lester et al. 2006; Hehemann et al. 2010; Xu et al. 2007). In addition to informing our understanding of microbial evolution, predictive models of gene transfer are needed to effectively engineer the human microbiome because HGT facilitates rapid adaptation to drugs and other perturbations (Lester et al. 2006; Hehemann et al. 2010). Until now, however, a dearth of available genome sequences and appropriate analytical techniques have left an incomplete view of the forces that govern HGT (Lawrence and Hendrickson 2003).

Many previous efforts to explore these forces have highlighted the relationship between phylogeny and HGT (Thomas and Nielsen 2005; Gogarten, Doolittle, and Lawrence 2002; Mazodier and Davies 1991; Lawrence and Hendrickson 2003). Phylogeny is expected to strongly influence HGT because shared evolutionary history is associated with overlap in the host range of mobile elements (Mazodier and Davies 1991), establishing a mechanistic basis for the phylogenetic control of gene exchange. Meanwhile, upon transfer, selection favors the persistence of genes acquired from close relatives, because these genes have greater compatibility with native molecular machinery (Tuller et al. 2011; Jain, Rivera, and Lake 1999).

Geography might provide an alternative structure to HGT by restricting dispersal, as suggested by the geographically organized distribution of *Vibrio cholera* integrons (Boucher et al. 2011) and NDM-1 antibiotic resistance genes (Kumarasamy et al. 2010).

A third possibility is that ecological similarity shapes networks of gene exchange by selecting for the transfer and proliferation of adaptive traits or by increasing physical interactions among community members. Reports of enriched levels of HGT among

hyperthermophiles (Aravind et al. 1998) and spatially segregated exchange among *Shewanella* isolates (Caro-Quintero et al. 2011) offer suggestive glimpses of such an ecological structure. However, it has been difficult to determine whether ecology plays a broader role in HGT due to the limited availability of genomes from similar environments and because most previous work has ignored the distinction between recent transfers and ancient events. The inclusion of transfers from millions or billions of years in the past can obscure ecological structure, because historical niches may not reflect modern environmental associations.

To explore the effects of phylogeny, geography and ecology on HGT we use an evolutionary rate heuristic to identify recent transfers among thousands of microbial genomes. Our heuristic finds blocks of nearly identical DNA (> 500 nucleotides, > 99% identity) in distantly related genomes (< 97% 16S rRNA similarity). HGT is the best explanation for these observations because the highly conserved 16S rRNA gene evolves about 25 times more slowly than protein-coding synonymous sites (Ochman, Elwyn, and Moran 1999). As a result, vertically inherited orthologs in such divergent genomes are nearly saturated with mutations at synonymous sites (Ochman and Wilson 1987), in contrast to the almost perfect identity that we require. To avoid over-counting transfers, we cluster similar genomes and normalize against the number of possible comparisons.

We have confirmed that at least 98% of all HGT events identified with our approach include a predicted protein-coding gene, indicating that potentially problematic non-coding elements do not significantly affect our results. To further validate our HGT detection method, we use two phylogenetic inference methods to evaluate the evolutionary origins of putatively transferred sequences. Quartet mapping and a gene loss analysis each support 99% of identified HGTs (Supplemental Fig. 1.1).

As expected, a large fraction of observed transfers (27%) include at least one predicted mobile element, underscoring the importance of these genes in facilitating exchange. However, when we account for redundancies we find that mobile elements like plasmids (2%), phage (1%), and transposons (9%) reflect only a promiscuous minority of the 10,770 total unique proteins that we observe, while the majority of unique genes (87%) provide other functions.

Direct exchange between any two bacteria in our dataset is unlikely, both because we limit our analysis to distantly related bacteria and because strains were isolated from different human subjects or environments, often on different continents. An average pairwise distance of 7,000 kilometers separates bacteria engaging in HGT. Therefore, each observed HGT likely reflects two independent acquisitions from a shared pool of mobile DNA, followed by proliferation.

To quantitatively explore the connectivity of bacteria in the human microbiome relative to other environments, we compare gene transfer among the 1,183 human-associated bacteria and 1,052 non-human associated isolates from a broad range of aquatic, terrestrial, and host-associated environments across the world. Even after correcting for biased sampling of human-associated clades (see Methods), pairs of bacteria isolated from the human body are 25-fold more likely to share transferred DNA than pairs from other environments ($P = 3.0 \times 10^{-270}$, combined Mann-Whitney U test).

This enrichment in human-associated transfer may be caused by the prevalence of overlapping selective pressures in the tightly regulated, endothermic human host compared to diverse, non-human environments that experience significant temporal and spatial variation in selective pressures. Consistent with this hypothesis, when the environment is specified more precisely by focusing on human isolates from the same body site, we observe two-fold higher rates of transfer ($P = 9.9 \times 10^{-108}$, combined Mann-Whitney U test). Remarkably, among the most closely related isolates from the same body site, this corresponds to recent HGT among > 40% of comparisons. This elevated transfer among bacteria isolated from similar environments extends beyond the human body, with three-fold more HGT among bacteria isolated from the same non-human environment relative to isolates from different non-human environments ($P = 1.3 \times 10^{-31}$, combined Mann-Whitney U test).

However, an alternative explanation for these observations is that closely related bacteria colonize similar environments, creating an apparent ecological effect that is actually driven by shared evolutionary history. To control for such a phylogenetic effect, we plot observed HGT over a range of phylogenetic divergences, and find that the strong enrichment for exchange within similar environments (same host, same body site, same non-human environment) persists across all distances (Fig. 1.1).

In order to directly compare the relative contributions of phylogeny and ecology to the enrichment in human-associated transfer, we compute recent HGT among bacteria isolated from the human body (same ecology) and between these human-associated bacteria and all non-human associated isolates (different ecology) over a range of phylogenetic distances. As shown by the dashed line in Fig. 1.2a, even the most deeply divergent bacteria that are separated by billions of years of evolution but share the same ecology, engage in more HGT than the mostly closely related isolates with different ecology. Thus, this recent gene exchange is structured by ecology more than by phylogeny.

We use a similar approach to explore the influence of geography relative to phylogeny, and find that exchange between continents is slightly lower than exchange within the same continent (Fig. 1.2b; $P = 0.02$, combined Mann-Whitney U test). However, this geographic effect is much weaker than that of phylogeny, which is itself less informative than ecology. Taken together, these analyses indicate that recent HGT frequently crosses continents and the Tree of Life to globally connect the human microbiome in an ecologically structured network.

This ecological architecture might only reflect the especially pronounced ecological differences between human-associated and non-human associated bacteria. To determine whether ecology has a broad influence on recent gene exchange we search for enriched HGT in narrower spatial, functional, and niche resolutions within the human host. Across all of these dimensions ecology strongly predicts gene exchange.

In addition to the previously discussed finding that transfer is enriched among bacteria from the same body site (Fig. 1.1), we find that further specifying the sub-site of isolation (e.g. separating vaginal isolates from other urogenital isolates) reveals even higher levels of transfer across all three annotated body sub-sites (sub-sites: vagina, gingiva, nasopharynx. Fig. 1.3a, Supplemental Figs. 1.2 and 1,3; $P = 1.7 \times 10^{-9}$, combined Mann-Whitney U test). When all human and non-human environments are considered, with scales ranging from tissues to ecosystems, we find that exchange at a narrow spatial scale, within an environment, always exceeds exchange at a broader spatial scale, with all other environments (Fig. 1.3b; $P = 1.3 \times 10^{-273}$, combined Chi-Square).

Until now, our analysis has relied on isolation environment as a proxy for ecological similarity, ignoring heterogeneities within these sites. Here we explore these differences, by evaluating the effects on HGT of oxygen tolerance and pathogenicity - the only other sufficiently annotated ecological features. Even after controlling for the effects of body site and phylogeny, we find that HGT is also structured by oxygen tolerance (Fig. 1.4a; $P = 7.7 \times 10^{-13}$, Chi-Square) and pathogenicity (Fig. 1.4b; $P = 7.4 \times 10^{-11}$, Chi-Square). These findings demonstrate that in addition to the extensive spatial effects described earlier, chemical gradients and symbiotic relationships provide further ecological structure to recent HGT. Because these results persist after controlling for explicit spatial effects, they appear to reflect selection rather than simply co-occurrence.

To further explore the role of selection, we probe its effects on the proliferation of different functional classes. If selection influences the rates and bounds of gene exchange, then the transfer of genes providing a non-specific selective advantage, like antibiotic resistance, should exhibit reduced environmental specificity relative to other, more niche-specific functional classes. To test this prediction, for each environment, we consider the fraction of observed transfers that include at least one antibiotic resistance gene (Fig. 1.3c). In contrast to our earlier observation of increased transfer within sites when all functional classes are grouped together (Fig. 1.3a and 1.3b), here we observe that resistance comprises a higher fraction of transfers across different environments than within the same environment (Fig. 1.3d; $P = 6.9 \times 10^{-279}$, combined Chi-Square). Thus, when ecological forces transcend environmental boundaries, mobile genes do too.

We have explored networks of gene transfer to evaluate the forces that influence recent HGT, finding that ecology is profoundly important. Now we demonstrate how knowledge of this association between ecology and HGT can be used to reveal clinical insights from patterns of observed gene transfer.

Our findings coupled with previous results (Hehemann et al. 2010) suggest that recently transferred genes among bacteria occupying a well-defined niche are especially likely to reflect adaptation to that niche. Consistent with this expectation, we find that many genes transferred among distantly related meningitis isolates - like hemolysins, adhesins, and antibiotic resistance genes (Supplemental Table 1) - are known to play an important role in the disease (Kim 2003). We suggest that other transferred genes with

unknown functions are likely cryptic virulence factors and should be prioritized for experimental annotation. Thus, in addition to recovering known virulence factors, our approach might streamline the search for novel drug targets (Clatworthy, Pierson, and Hung 2007), because while it is prohibitively difficult to explore all 24,095 unique meningitis genes with unknown function, it is tractable to evaluate the thirteen that were recently transferred. We use this approach to identify genes associated with other diseases (e.g. pneumonia, endocarditis; Supplemental Tables 2 and 3) and environments (e.g. hot springs and soil; Supplemental Tables 4 and 5) opening a molecular window into the genetic traits that define ecological niches.

As a second example, our analysis of recent HGT reveals potential sources of clinical antibiotic resistance. We find that bacteria from farm animals and human food are enriched in transfer of resistance with human-associated bacteria relative to other non-human associated isolates ($P = 1.7 \times 10^{-11}$ and $P = .01$, respectively, Mann-Whitney U test). Forty-two unique antibiotic resistance genes are transferred between human and farm isolates. These transferred genes comprise nine families, all of which include both genes known to provide resistance to clinical antibiotics and genes known to confer resistance to agricultural drugs (see Supplemental Table 6). This suggests that livestock-associated bacteria can contribute to clinical resistance without directly infecting humans, because for these mobile traits, genes, not genomes serve as the unit of evolution and proliferation. Moreover, we observe forty-three unique antibiotic resistance genes crossing national borders, suggesting that because the human microbiome is globally connected, local contamination of the shared mobile gene pool can have significant trans-national consequences.

Here we present the discovery that ecology governs recent HGT and use this finding to reveal the key genes and networks of exchange that facilitate colonization, and occasionally exploitation, of the human host. In the future this approach could be extended to analyze bacterial genomes from individuals or groups of individuals that differ in diet, disease, or descent to search for the microbial genes that affect these human conditions.

1.3 Methods

Methods Summary

All 16S rRNA genes were identified using the GreenGenes database (DeSantis et al. 2006a). 115 genomes with spurious or truncated 16S rRNA sequences were excluded from our analysis. We used BLAST (version 2.2.20) with default parameters (Altschul 1990) to calculate an all against all nucleotide alignment for 2,235 genomes downloaded from IMG (Markowitz 2006). We infer HGT events from blocks of nearly identical DNA (> 99% identity, > 500 bp) in distantly related genomes (< 97% 16S rRNA similarity). To avoid over-counting events in ancestral lineages, we collapse closely related genomes using average linkage clustering into groups ('species') with 16S rRNA dissimilarity of 2%. For each pair of these clusters, we calculated the fraction of genome comparisons between clusters that share at least one inferred HGT event. We sum this fraction over all pairs of clusters and normalize to the total number of comparisons in order to calculate the HGT per 100 comparisons. Statistical tests of HGT enrichment were performed separately for each distance bin then combined into a single p-value using Fisher's Method. We modeled antibiotic resistance transfer as a binomial random variable with parameter p and calculated a 95% confidence interval around our estimate of p. The size of this confidence interval, which is the statistical uncertainty of our estimate, was used to desaturate the color of the heatmap in Fig. 1.3c. To explore the effects of oxygen tolerance and pathogenicity on HGT, we use a Chi-Square test to compare the observed frequency of HGT to the expected value given the distribution of body sites and phylogenetic divergences. Protein-coding regions were identified and annotated using BLASTX (Altschul 1990) (E-value < 1E-50) and UBLAST (R. C. Edgar 2010) (maxtargets = 100, E-value < 1E-50) searches against the NCBI nr database. Unique genes reflect unique best BLAST hits to the database. Antibiotic resistance genes were annotated using the Antibiotic Resistance Genes Database (Liu and Pop 2009).

Extended Methods

Quartet mapping

To test whether phylogenetic reconstruction supports our inference of HGT, we performed quartet mapping, in which all possible four-member trees are generated and analyzed to simulate analysis of the larger and more computationally challenging parent tree. We followed a similar approach to the quartet mapping described by Daubin and Ochman (Daubin and Ochman 2004). Briefly, we searched all 2,235 genomes in our analysis for homologs to each HGT event (defined as best reciprocal BLAST hits with > 60% nucleotide identity over > 60% of the length of the transferred gene; see note on homology below). For HGT events with at least two homologs, we used MUSCLE (with default settings) to construct an alignment of the HGT sequences and all other non-HGT sequences. Events with fewer than two non-HGT homologs - 23% of the total - cannot be used to generate a quartet and so could not be analyzed by quartet mapping. For the quartets that remained, we used Tree Puzzle to analyze all possible quartet topologies among the aligned HGT and non-HGT sequences. With Tree Puzzle we used exact parameter estimates and gamma distributed rates with four rate categories. To provide phylogenetic confirmation of our putative HGT events, we computed the likelihood of obtaining a quartet grouping the HGT events together, versus the alternative, vertical model that would group sequences by the topology of the species phylogeny. A previously published likelihood ratio (Daubin and Ochman 2004) was then used to place phylogenetic confidence in each HGT event. We used the most stringent confidence threshold possible, requiring a likelihood ratio of 1.0 to support HGT inference. With this conservative approach, more than 99% of the HGTs we analyzed were supported.

Gene loss analysis

We explored whether vertical inheritance is a plausible alternative explanation for each inferred HGT by determining the minimum number of independent loss events that would be needed to support a model of vertical inheritance. We mapped all inferred transfers and their homologs to the IMG species tree and calculated the number of independent loss events that would be required to explain the sparse phylogenetic distribution of these events. Here, we define homologs as best BLAST hits with > 90%

identity and > 80% length (see note below). These parameters allow for considerable variation in evolutionary rates within the gene family.

As shown in Supplemental Fig. 1.1, for the majority of HGT events, over 100 independent loss events would be required to accept a model of vertical descent. To contextualize this remarkable observation, most parsimony based HGT detection tools use an empirically derived estimate of approximately 3:1 as the parsimony cost of losses relative to HGT (David and Alm 2011). Using this 3:1 parsimony metric, over 99% of our events can be explained by HGT.

Note on the detection of homologs

We varied the parameters that define homology for the two approaches above in order to maximize our ability to detect vertical transmission. We used an especially permissive definition of homology for quartet mapping to allow a maximal number of potentially homologous genes to disrupt the pairing of the putatively transferred sequences, thereby increasing the opportunity to return a quartet that does not support HGT. We employed a more moderate definition of homology for the loss analysis to avoid spuriously inserting unrelated proteins that may have appeared as false loss events.

Controlling for contamination

To control for the potential effect of contamination derived from genomes processed at the same sequencing facility, we repeated our principal analysis, but only compared genomes sequenced at different facilities. This restricted analysis confirmed that our main findings are not caused by contamination between projects at the same sequencing center. In Supplemental Fig. 1.4, we show that there is more HGT among human-associated bacteria than among non-human associated bacteria, across all phylogenetic distances. The enrichment in HGT among bacteria occupying the same body site relative to bacteria occupying different body sites is similarly replicated in this restricted analysis (as found in Figure 1.1 of the main text).

In Supplemental Fig. 1.5, we also show that the most distantly related comparisons with

shared ecology continue to exchange more DNA than the most closely related comparisons with different ecology when only HGT between sequencing centers is allowed (as found in Figure 1.2 of the main text).

Controlling for cosmopolitan genomes

To control for the potential effect of cosmopolitan genomes that inhabit multiple environments, we repeated our principal analysis, excluding all genome clusters containing at least two representatives from different body sites, hosts, or other environmental categories. This removed cosmopolitan groups of organisms like *Escherichia coli*, which is found in the gut, skin, blood, and non-human environments for example. This restricted analysis robustly yields the pattern of ecological enrichment found in the main text (Supplemental Fig. 1.6).

Limitation of HGT detection

Our method is only able to detect horizontal gene transfer between distantly related lineages. Another limitation is that our method can only detect recent events that share 99% nucleotide identity. Consequently the dynamics discussed in our analysis may not apply to more ancient HGT or to HGT between less divergent strains. However, because a stringent phylogenetic distance cut-off is used to inform each HGT classification our method avoids many of the limitations of previous BLAST-based approaches to HGT detection (Stanhope et al. 2001).

Limitations of geographic inference

There are a few important caveats to consider when reviewing our geographic findings. First, due to limited sample size, we only explored the effects of geography at continental scales. It is possible that strong effects may persist at finer spatial scales, although these may be primarily driven by ecological overlap, which is difficult to distinguish from local geography. Second, the location of isolation is only a proxy for the overall geographic range of a sequenced strain. When a strain is isolated from a particular site, it may have a range that extends across a much larger geographic range, obscuring the validity of geographic inference from a single sample.

Annotation of mobile genetic elements

For this analysis we were interested in exploring the approximate magnitude of mobile elements relative to other functional groups. In the interests of defining the minimum number of mobile elements in our analysis, we chose a rapid and highly specific method at the expense of sensitivity. We aligned all transferred sequences to the NCBI nr database using BLASTX. We extracted the annotations for the best BLASTX hit in nr (with an e-value of e < 1E-50). Next we used keyword search text mining coupled with manual curation to count the frequency of each functional category. Our keywords are designed to reduce false positives - we understand that valid mobile elements may not be detected with this simple approach.

The keywords used to identify each functional group are listed below (case sensitive):

Transposons: transpos*, TN, insertion element, is element, IS element
Phage: phage, tail protein, tegument, capsid
Plasmid: relaxase, conjugal transfer, Trb, relaxosome, Type IV secretion, conjugation, Tra[A-Z], Mob[A-Z], Vir[A-Z][0-9], t4ss, T4SS, resolvase
Other MGE: recombinase, integrase

The percent of total proteins (27%) is calculated by counting each of the functional classes as a fraction of all transferred sequences. In order to account for redundancy in the set, we extract the NCBI gene identifier for the best BLASTX hit for each transferred sequence. We then remove all redundancies from this list of gene identifiers and count the fraction of unique gene identifiers that fall into each of the functional classes described. This analysis suggests that a relatively small group of promiscuous mobile elements accounts for a large fraction of total transferred sequences.

Definition of environments

Farm samples are taken directly from animals used in agriculture (horse, cow, sheep, goat, pig). As with human subjects, samples from animals vary (blood, stool, rumen etc.).

Metadata to define environments, such as isolation site, oxygen tolerance, and pathogenicity were downloaded from IMG[27].

Treatment of ambiguous metadata annotations

We only consider genome comparisons for which we have appropriate metadata. For genomes with partial metadata (i.e. oxygen tolerance is annotated, but continent and disease are missing), we include the genome when possible (for oxygen tolerance) and ignore it in other analyses (continent and disease).

When comparing the frequency of HGT in the same environment with the frequency of HGT between different environments it is necessary to handle ambiguous genome annotations with multiple annotated environments (e.g. gut and skin). In these cases, we consider this strain once for each metadata label. Thus when a strain from the gut is compared to a strain annotated as gut and skin, this comparison will contribute to both comparisons of gut-gut transfer and gut-skin transfer.

Computation of error bars

Error bars reflect our estimated uncertainty in the sampling of a binomial random variable (the observation of HGT). We compute error bars as the standard deviation in %HGT by modeling the total number of transfers as a binomial random variable with parameters p and n. We take n to be the number of independent species cluster comparisons and we estimate p as the total %HGT observed at each phylogenetic distance. From these considerations, it follows that the variance is given by $\mathrm{Var}[\%\mathrm{HGT}] = p(1-p)/n$ which is used to calculate the standard deviation at each distance bin.

Counting HGT

When measuring the frequency of HGT between environments we only consider the fraction of genomes that share at least one HGT. We do not consider the length of a transfer because high variance in event length would add significant noise to our results and overweight rare, large transfer events that do not reflect evolutionary independence. We do not consider the number of distinct regions of HGT shared between two genomes

because transposition or poor assembly might falsely inflate this metric by splitting a single large event into many smaller apparent events.

In Fig. 1.3 of the main text, HGT is computed as the average across all distance bins in contrast to Fig. 1.1, where HGT is computed in separate distance bins. As a result, the frequencies of HGT cannot be directly compared between the two figures.

Clustering similar genomes

In order to avoid over-counting transfers, we use average linkage clustering to group similar genomes (with < 2% average 16S rRNA divergence). This ensures that transfers between clusters reflect evolutionary independence and avoids the problem of counting a single transfer in a densely sampled lineage many times. All comparisons discussed in the text reflect transfers across clusters constructed in this manner.

Because the sequenced flexible genome is larger when more isolates from a single cluster are considered, the probability of observing at least one transfer between two clusters with many sequenced isolates is greater than between two clusters with fewer sequenced isolates. To account for this effect, for each cluster comparison we consider the fraction of genomes that share an HGT. We equally weight all genome comparisons between two clusters. If 50% of a genome cluster has a hit with at least one member of another genome cluster, we consider this cluster comparison as 50% of an HGT.

Statistical methods

To test for overall enrichment in HGT between two metadata labels (e.g. human vs. non-human) we perform separate statistical tests for enrichment within each phylogenetic distance bin, then combine these test results into a single p-value using Fisher's method. Within each phylogenetic distance bin, we determine if there is a significant difference in HGT frequencies between all pairs of genome clusters belonging to the two different metadata labels. With our counting and clustering protocols (described above), we create two vectors (each corresponding to a metadata label) of HGT frequencies (with continuous values) that we compare with a Mann-Whitney U-test. This approach is

applied to assess differences in observed frequencies of HGT and to assess the statistical significance of the data underlying Fig. 1.1, Fig. 1.2, and Fig. 1.3 in the main text. This approach controls for the effect of phylogeny by restricting comparisons of HGT frequency to isolates of similar phylogenetic divergences (distance bins of 1% 16S rRNA distance).

After establishing the strong effect of body-site on HGT frequency in the human microbiome, further analyses (such as oxygen tolerance and pathogenicity as in Fig. 1.4, main text) must control for both the effects of phylogeny and body-site. We achieve this by calculating the frequency of HGT for all possible combinations of body-sites and phylogenetic divergences. For example, the expected value for skin-gut transfer at 3-4% 16S rRNA divergence is the average of all observations that meet these metadata criteria. Our null model assumes that further constraining our analysis with additional metadata labels will not lead to values that deviate from these expected values. To test this model, we compare the expected value to the observed frequency of HGT when the analysis is further conditioned on a new metadata label (e.g. anaerobes in skin and gut at 3-4% 16S rRNA divergence). We determine whether this further metadata constraint is associated with elevated HGT by using a Chi-Square test to compare the expected values with the observed values.

1.4 Figures

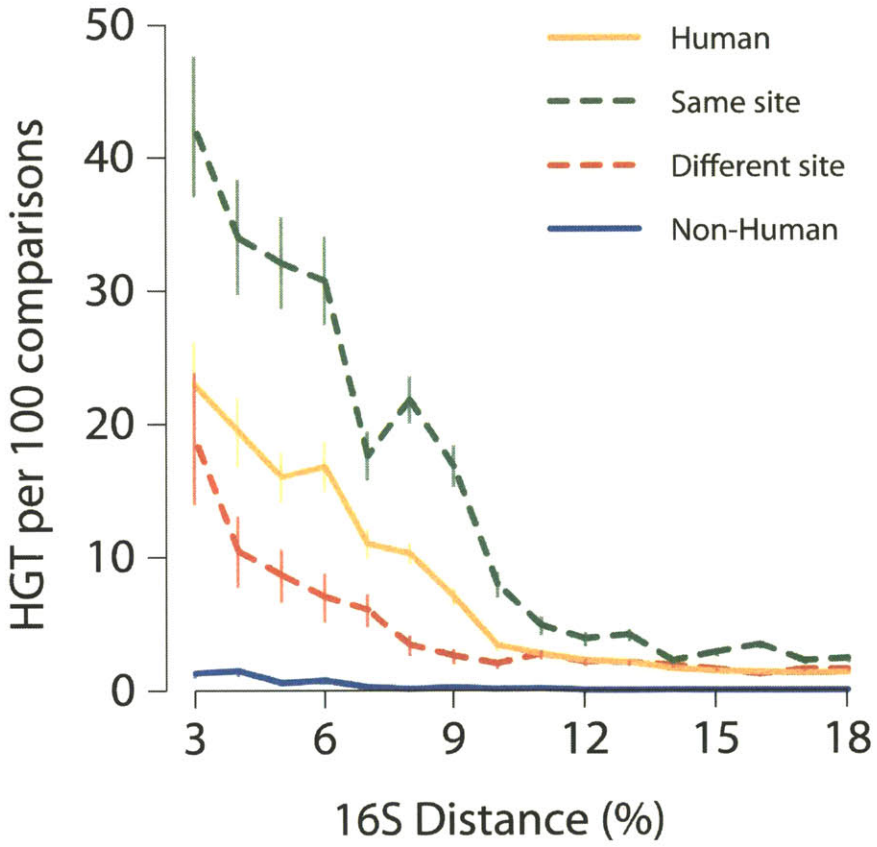


Figure 1.1: Recent HGT is enriched in the human microbiome across all phylogenetic distances. These plots (**a**, **b**) show HGT frequency as a function of the phylogenetic divergence between species, for **a**, human-associated bacteria, and **b**, non-human associated bacteria. We define species as clusters of genomes separated by < 2% 16S rRNA divergence. HGT frequency is calculated in bins of 1% 16S rRNA divergence. Error bars reflect one standard deviation (see Supplemental Methods), with sample sizes described in Supplemental Table 8. These trends are also observed after controlling for the potential effects of sequencing center contamination (Supplemental Fig. 1.4) and cosmopolitan strains (Supplemental Fig. 1.6).

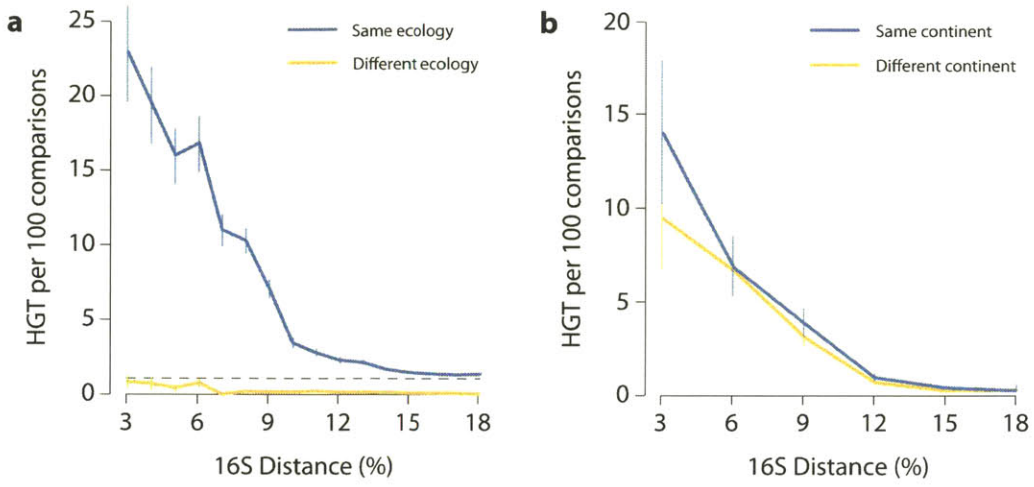


Figure 1.2: **Ecology is the dominant force shaping recent HGT in the human microbiome. a**, The frequency of HGT among human-associated isolates (same ecology; blue) and between human-associated and non-human associated isolates (different ecology; red). **b**, The frequency of HGT among bacteria isolated from the same continent (blue) and different continents (red). Due to reduced sample size in **b**, we pooled comparisons into larger phylogenetic distance bins of 3%. Error bars are calculated as in Fig. 1.1. The role of ecology in (**a**) is recovered when we control for sequencing center contamination (see Supplemental Fig. 1.5).

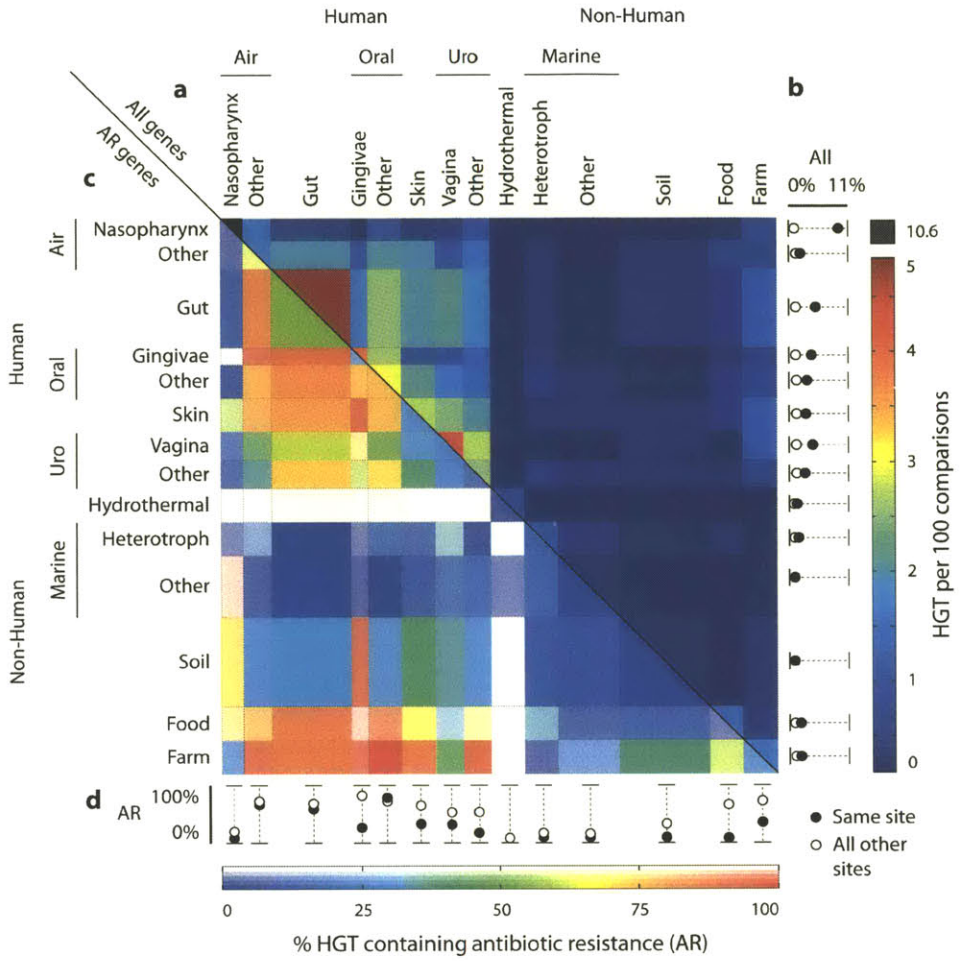


Figure 1.3: HGT is ecologically structured by functional class and at multiple spatial scales. The frequency of transfer among different environments is shown for all functional groups (**a**, **b**) and for antibiotic resistance (AR) genes only (**c**, **d**). Box widths indicate the number of genomes from each environment. **a,** When all genes are considered (upper half) human isolates form a block of enrichment (upper left). **b,** For every environment examined we observe more transfer within the same environment (black dots) than between environments (white dots). **c,** The fraction of gene transfers that includes at least one AR gene for each environment. Statistical uncertainty in the proportion of AR transfer is indicated by reduced color saturation (see Methods). **d,** AR genes comprise a significantly higher fraction of observed HGT between different environments (white dots) relative to within the same environment (black dots) in contrast to (**b**).

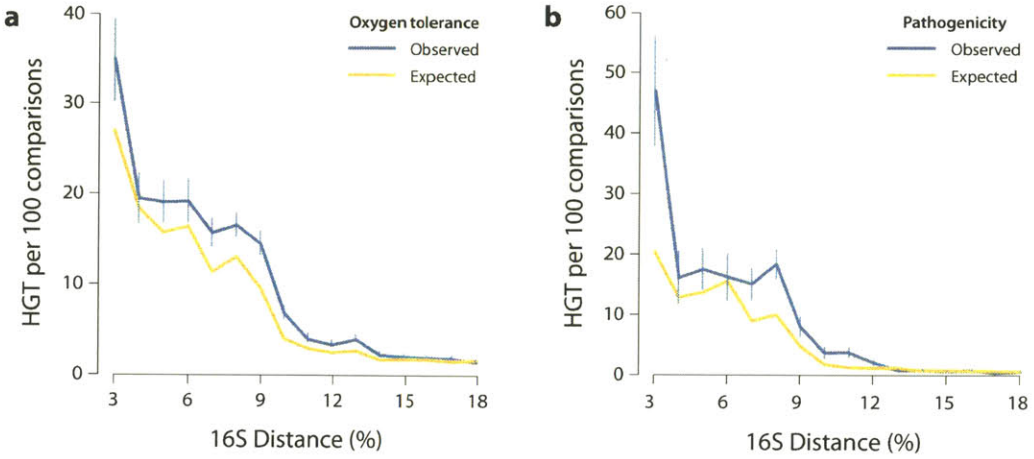
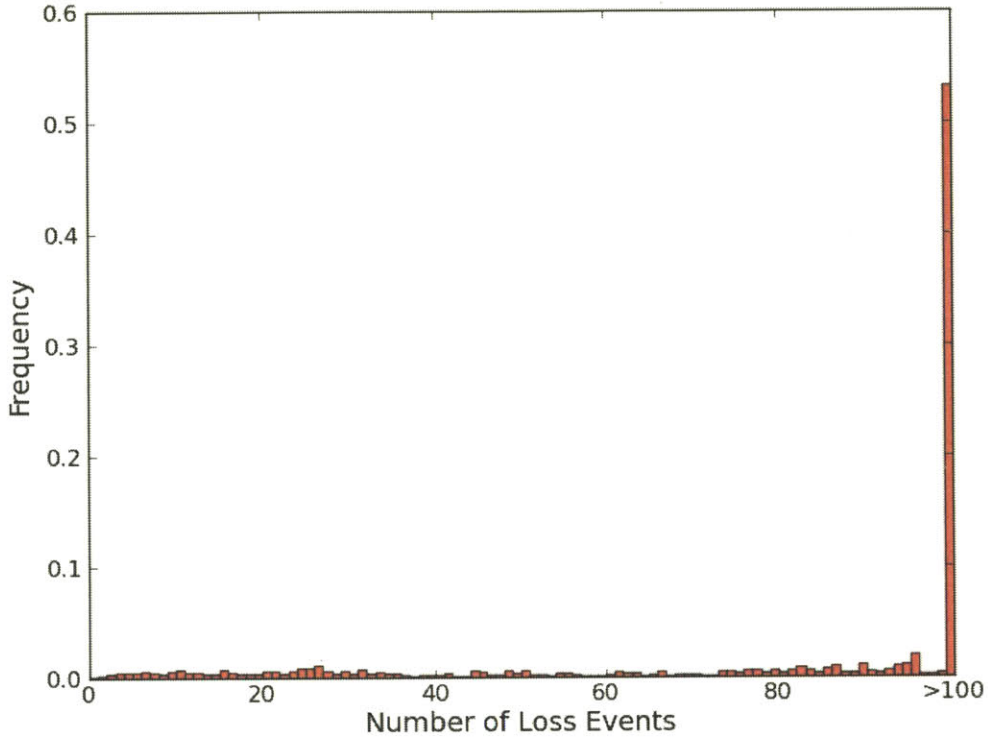
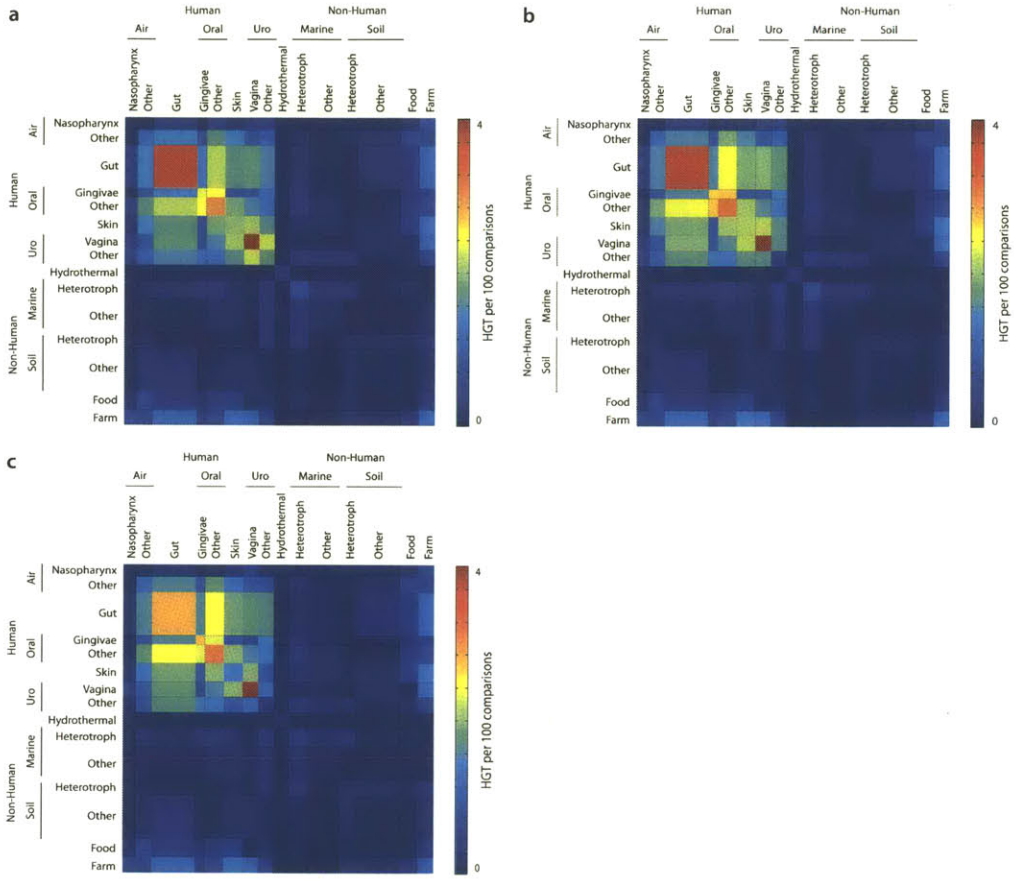


Figure 1.4: Gene exchange is ecologically structured by oxygen tolerance and pathogenicity. The frequency of HGT between genomes with the same (**a**) oxygen tolerance and (**b**) pathogenicity is shown relative to their expected values. Expected values are based on overall frequencies of transfer among bacteria from the same distribution of body sites and phylogenetic distances. Bacteria that share the same oxygen tolerance (aerobic, anaerobic, microaerophilic, or facultative aerobic) and pathogenicity (pathogenic or commensal) engage in significantly more HGT than is expected under the null model. Error bars are calculated as in Fig. 1.1.

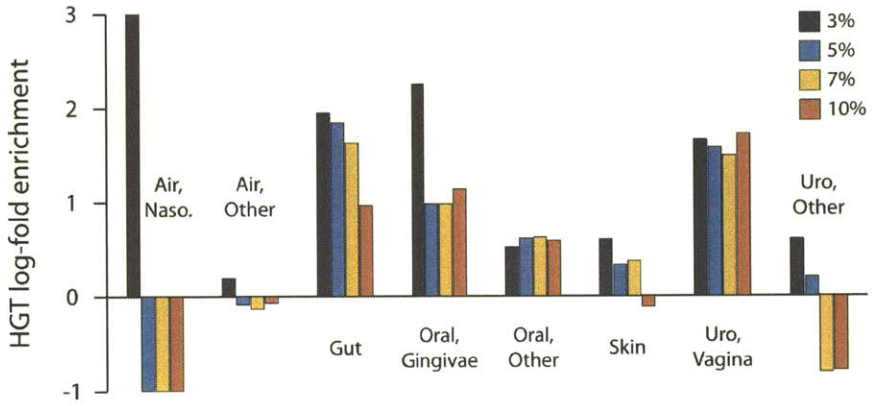
1.5 Supplemental Figures and Tables



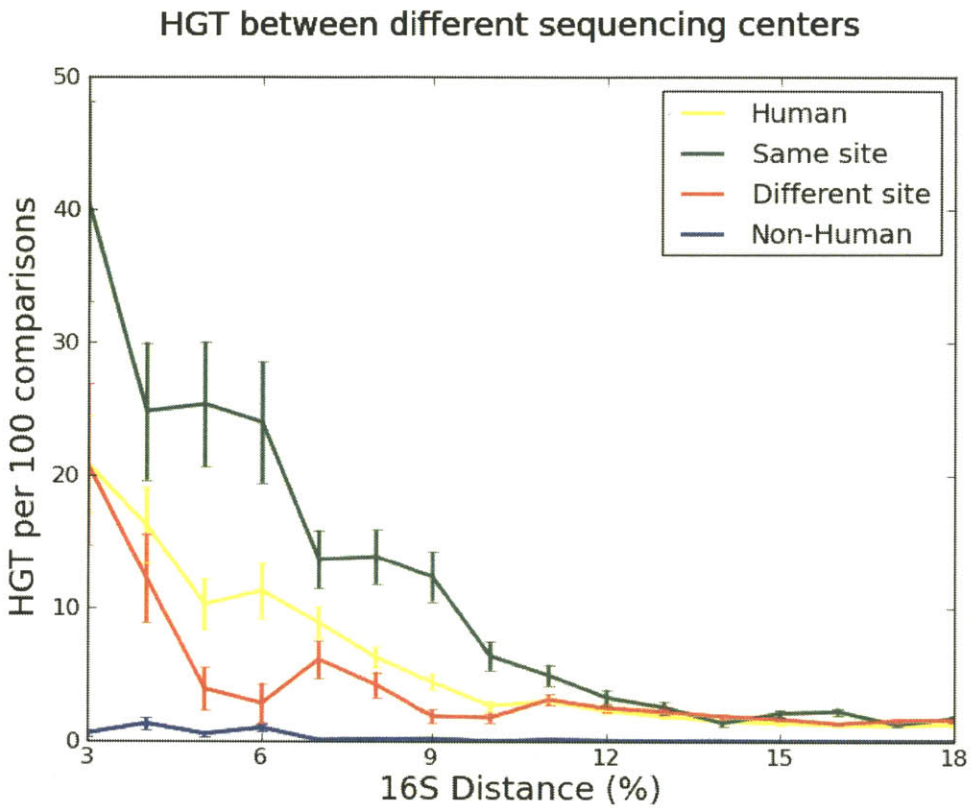
Supplemental Fig. 1.1: The majority of inferred HGT events require over 100 independent loss events in order to accept a model of vertical descent. For each inferred transfer we map homologs onto the species tree and infer the minimum number of independent loss events needed to support a model of vertical inheritance. This figure depicts the frequency with which loss events are inferred – most inferred transfers would require extensive loss events in order to accept the alternative model of vertical transmission, supporting our approach to HGT detection.



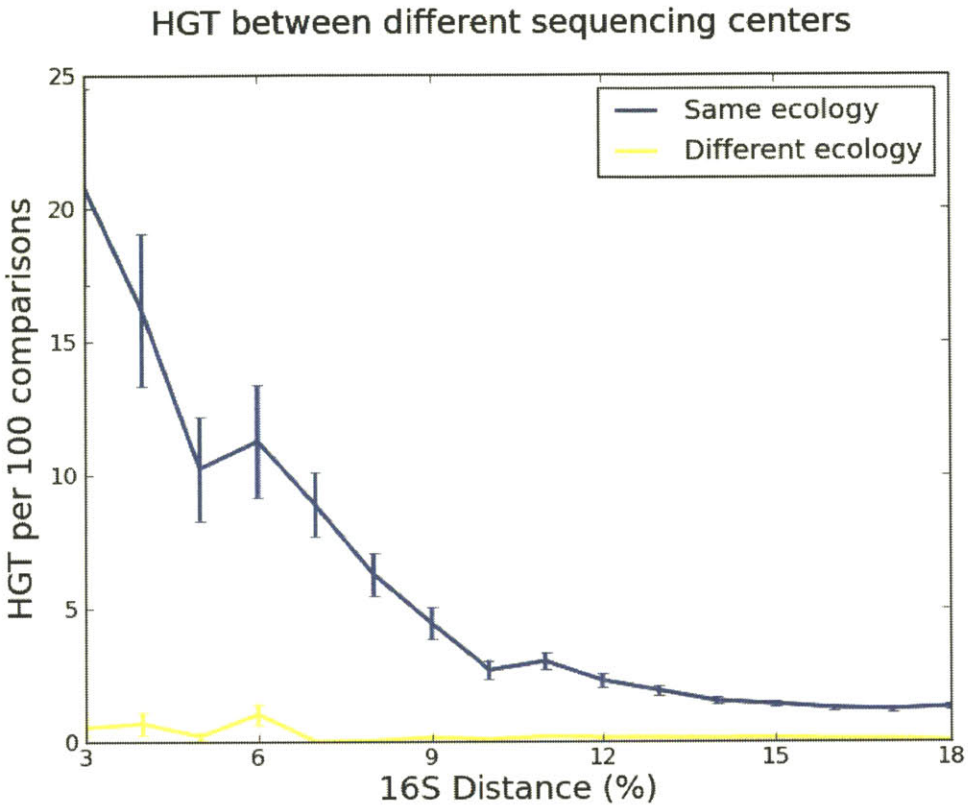
Supplemental Fig. 1.2: Heatmap of HGT among isolates in different environments at 5%, 7% and 10% 16S rRNA divergences. This figure shows the frequency of HGT between each of the environments included in this study across three different distance cutoffs, in addition to the overall plot shown in Figure 1.3a of the main text. Each distance cutoff includes all comparisons satisfying the given separation criteria (e.g. 5% includes comparisons of all clusters of bacteria separated by at least 5% 16S rRNA divergence). Although the specific values of enrichment vary across different distance cutoffs, the overall pattern of human, body site and body sub-site enrichment persists across all distance groupings. We show only the heatmap for all gene classes (excluding the inset heatplot for antibiotics that appears in the main text Figure 1.3c) because there are insufficient counts to yield reliable estimates for rates of long distance transfer when only antibiotic resistance genes are considered.



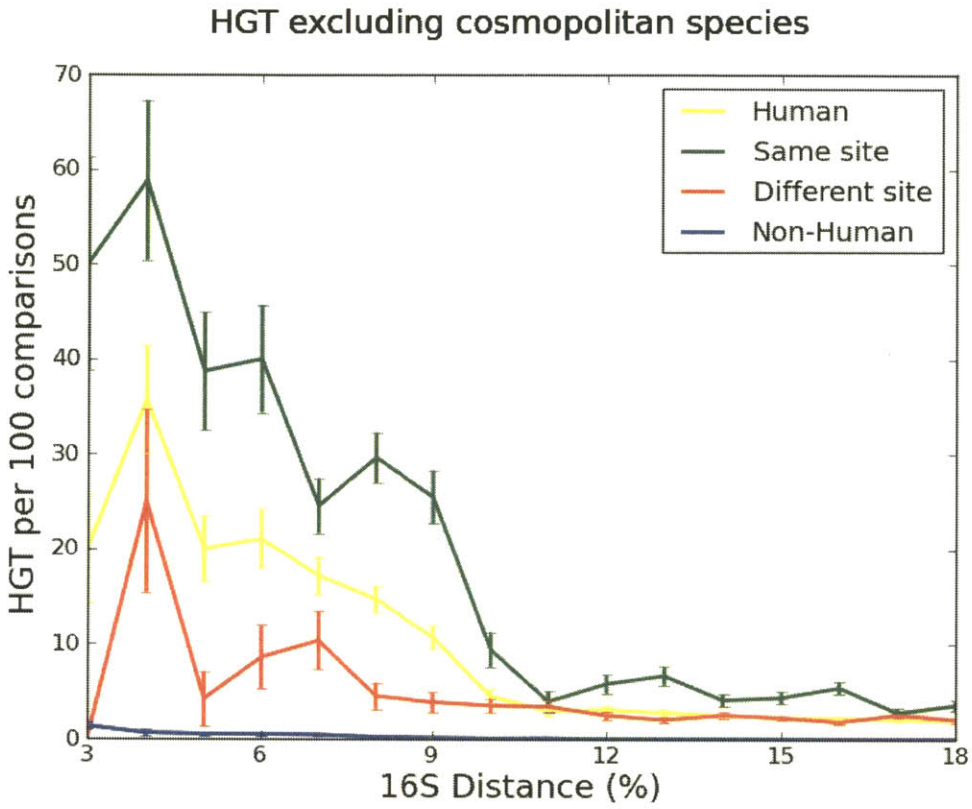
Supplemental Fig. 1.3: Barplot of HGT for each body site at 3%, 5%, 7% and 10% distance cutoffs. This figure summarizes the persistence of body-site and sub-site enrichment across four distance cutoffs. As in Supplemental Fig. 1.2, distance cutoffs reflect all comparisons with at least the given 16S rRNA distance. The log-fold enrichment indicated on the vertical axis describes the ratio of observed transfers within the given body site at each distance relative to HGT among all human isolates at the same phylogenetic distance cutoff. The poorly sampled nasopharynx (n = 25) and non-vaginal urogenital sites (n = 46) are the only categories for which the enrichment in transfer does not persist across phylogenetic distances (likely due to uncertainty arising from small sample sizes). Otherwise, the majority (n = 480) of isolates belong to body sites for which enrichment persists across all observed distances.



Supplemental Fig. 1.4: Ecological structure persists when only genome comparisons from different sequencing centers are allowed. We compute the frequency of transfer within human associated isolates (yellow), non-human isolates (blue) human isolates from the same body site (green) and human isolates from different body sites (red), while only allowing genome comparisons between different genome sequencing centers. This controls for contamination that might arise in the sequencing and assembly process.



Supplemental Fig. 1.5: Ecology is the dominant force shaping recent HGT in the human microbiome, even when HGT is only allowed between different sequencing centers. This figure compares the effects of ecology relative to phylogeny on HGT, when HGT is only allowed between different sequencing centers. The frequency of HGT is shown among human-associated isolates (same ecology, blue) and between human-associated and non human-associated isolates (different ecology, yellow). Even the most distantly related bacteria with shared ecology engage in more HGT than the most closely related bacteria with different ecology when we control for contamination caused by sequencing projects from the same sequencing center.



Supplemental Fig. 1.6: Ecological structure persists when cosmopolitan species are excluded. We compute the frequency of transfer within human associated isolates (yellow), non-human isolates (blue) human isolates from the same body site (green) and human isolates from different body sites (red), while excluding species that are present in multiple environments (cosmopolitan species). This controls for the potential confounding effect of cosmopolitan species.

Supplemental Table 1.1: Sample sizes used in statistical comparisons. This table shows the sample sizes used in the Mann-Whitney U-tests in Figs. 1.1, 1.2, and 1.4.

	16S rRNA Distance Bins								
Environment	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Human	166	232	378	383	827	1327	1638	2672	3544
Human within	88	120	184	193	416	546	562	786	973
Human between	62	132	200	198	372	576	857	1689	2300
Non-Human	1658	1169	2859	2657	3810	6526	6891	10380	13841
Same ecology	166	232	378	383	827	1327	1638	2672	3544
Different ecology	552	425	948	911	1876	3831	3979	6990	8263
Same continent	84			253			652		
Different continent	108			372			1186		
Same oxygen tolerance	106	190	268	247	500	708	683	988	1297
Same pathogenicity	30	70	122	84	174	233	266	355	434

	16S rRNA Distance Bins						
Environment	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Human	5001	7262	10802	15319	18587	18125	14944
Human within	1417	1896	2807	3935	4556	4275	3451
Human between	3101	4704	6840	9586	11739	10982	8946
Non-Human	22295	29688	45310	64169	76375	74374	58819
Same ecology	5001	7262	10802	15319	18587	18125	14944
Different ecology	13433	18909	27956	39740	50099	50868	40880
Same continent	2120			4709			1974
Different continent	3375			7831			3510

Note: Supplemental Tables 1-6 from this publication can be found online at nature.com. These tables are very long and have been omitted from this document due to space constraints.

Chapter 2: Identical CRISPR spacers among distantly related bacteria reveal common strategies to target promiscuous mobile elements.

Smith MB, Alm, EJ (2014) Identical CRISPR spacers among distantly related bacteria reveal common strategies to target promiscuous mobile elements. (*In progress*)

2.1 Abstract

The discovery of ubiquitous, recent horizontal gene transfer (HGT) in bacteria suggests that experimental observations may understate the host range of mobile genetic elements (MGE) (Smillie et al. 2011). As a form of heritable, adaptive immunity in bacteria, arrays of Clustered Regularly Interspersed Short Palindromic Repeats (CRISPR) provide a natural record of infection history that can be used to systematically explore the host range of MGE. Through the analysis of 159,468 CRISPR spacers in 3,314 bacterial genomes, here we show that even distantly related bacteria with less than 90% homology at the 16S rRNA gene often share matching CRISPR spacers with identical sequences (7,009 observations), far more than expected by chance ($P < 1 \times 10^{-200}$, Chi-Square). Shared spacers are more likely to share homology to sequenced phage (10%) or plasmid genomes (2.7%) than are unique spacers (2.3% and 0.8% respectively), suggesting that shared spacers target common elements in MGE and may confer broad resistance. At least 52% of bacteria with matching spacers also share identical CRISPR repeats, indicating that entire CRISPR arrays are often horizontally transferred. These observations imply that targeted elements are widely shared and that bacteria recycle a surprisingly narrow set of effective molecular strategies to target MGE.

2.2 Main Text

Mobile genetic elements like phage and plasmids link bacterial genomes through horizontal gene transfer. Although some plasmids are understood to have a very broad host range, even spanning gram positive and negative bacteria (Rawlings and Tietze 2001), the frequency and implications of this capability have not been systematically evaluated. Moreover, despite intriguing anecdotes to the contrary (Chen and Novick 2009; Lester et al. 2006; Rawlings and Tietze 2001), most phage are believed to maintain relatively narrow host ranges (Hyman and Abedon 2010). Yet this understanding of MGE host-range has emerged from a limited view of evolution constrained by the restrictions of traditional culture-based methods for characterizing interactions between bacteria and the mobile elements that pass between them. Due to the difficulty of exhaustively culturing all pairs of hosts and MGE, nearly all efforts have focused on closely related organisms. Long-distance relationships between mobile elements and distantly related bacteria have not been systematically probed (Weitz et al. 2013). The implicit assumption is that because many mobile elements show a limited host-range among closely related bacteria, the same elements are unlikely to infect even more distantly related strains (Hyman and Abedon 2010).

However, this reasonable premise overlooks the vast combinatorial matrix of interactions that characterizes the microbial world. With a trillion cells in a gram of stool, even rare events are expected to happen frequently (Ott et al. 2004). Current opinion regarding MGE host range is heavily influenced by culture-based methods, which are poorly suited to the identification of relatively rare events. Moreover, an event need only happen once over a lineage's history to impact its evolution.

The recent emergence of low-cost sequencing technology has enabled the sequencing of thousands of bacterial genomes and the phage and plasmids that are associated with them. Comparative genomic analyses have demonstrated that HGT is far more common than previously appreciated, and that such transfers occur across great evolutionary distances, often over very short evolutionary time-scales (Smillie et al. 2011). These observations

provide compelling evidence to revisit previous assumptions about the phylogenetic reach of mobile genetic elements.

In addition to HGT, CRISPR arrays provide another, more direct resource for assessing the host range of mobile genetic elements that does not require exhaustive experimental characterization. CRISPR provide bacteria with heritable, adaptive immunity against phage and other mobile genetic elements. Each array contains alternating sequences known as repeats, which are nearly identical throughout an array, and spacers, which are generally unique pieces of DNA complementary to a sequence in an MGE target. When expressed as an RNA transcript along with associated Cas genes, CRISPR spacers provide sequence-specific targeting of foreign DNA, identifying complementary sequences for degradation (van der Oost et al. 2009; Sorek, Kunin, and Hugenholtz 2008). Because spacers are integrated into the host genome, when a new spacer is added, it is passed along to daughter cells, providing heritability. As new spacers are integrated, old spacers are deleted. Consequently CRISPR arrays form a historical record of recent infections by mobile genetic elements that can be mined for new information about host range (Tyson and Banfield 2008). As a test for the range over which these interactions occur, here we seek to determine whether bacteria from different genera possess identical spacers that may target shared mobile genetic elements.

We consider all 9,472 bacterial genome sequences available from GenBank and use CRT to identify 12,811 CRISPR arrays among 4,886 of these available genomes (Bland et al. 2007). Among these genomes containing CRISPR, we restrict our analysis to the 3,314 genomes that have a full-length 16S rRNA gene to enable phylogenetic placement. We use pairwise alignments to identify the 16S rRNA dissimilarity among all pairs of genomes. Next we take each of the 159,468 spacers identified in these genomes and use BLAST to find pairs of spacers that share at least 30 base pairs with 100% identity (the mean length of discovered spacers was 35.1 base pairs). We do not require perfect identity over the entire sequence, because of considerable heterogeneity in spacer length. Surprisingly, we found 27,098 pairs of genomes sharing at least one identical spacer region.

Given that this analysis compares 10.9 million pairs of genomes, however, it is important to determine how often this observation is expected to occur by chance as a simple result of common sequence motifs. To evaluate this null hypothesis, for each spacer included in our analysis, we choose a random sequence of the same length from another location in the genome. We take these synthetic spacers and using the technique outlined above, identify pairs of genomes where these synthetic spacers happen to match other synthetic spacers. We find only 109 cases of matching synthetic spacers. Genuine spacers match much more often than synthetic spacers ($P < 1 \times 10^{-200}$, Chi-Square).

Phylogenetic history provides another confound that must be controlled to evaluate the significance of this observation. Many comparisons in this analysis include closely related genomes, where matching spacers are likely inherited vertically rather than through independent evolution. To control for this effect, in Figure 2.1, we plot the relationship between 16S rRNA divergence and the fraction of genome comparisons containing at least one matching spacer. Indeed, 22% of genomes with CRISPR that are less than 1% divergent at the 16S rRNA gene share an identical spacer, accounting for 18,915 (70%) of all genome pairs with matching spacers. In addition to these ubiquitous short-range interactions which are almost certainly due to vertical inheritance, we observe many long distance interactions. For example, we found 7,009 pairs of bacterial genomes that have at least one matching spacer and are at least 10% divergent at the 16S rRNA gene (versus 63 synthetic spacers, $P < 1 \times 10^{-200}$, Chi-Square). Identical sequences of this length are unexpected through vertical inheritance at such great phylogenetic distances.

These long-distance spacer matches must be explained through an alternative evolutionary mechanism. One interpretation is that these spacers are actually the product of HGT themselves. There are several distinct classes of CRISPR and each has a distinct repeat pattern that is generally phylogenetically restricted (Kunin, Sorek, and Hugenholtz 2007). As a result, these repeats can be used as a marker of the host of origin. If a matching spacer is the product of HGT, then the associated repeats should also be identical. Indeed, when we compare the repeats associated with matching CRISPR, we

find that 52% of repeats are identical. As shown in Figure 2.2, this pattern persists across all phylogenetic bins, strongly indicating horizontal transmission in these cases. Furthermore, when we consider pairs of genomes that have CRISPR, but do not share a matching spacer, we find that fewer than 0.6% of repeats are identical. And, as shown in Figure 2.2, nearly all of these identical repeats are among closely related comparisons. This very low rate of homology among CRISPRs without matching spacers supports the view that most CRISPR repeats are phylogenetically restricted. Many genomes with matching CRISPR spacers also contain matching repeats, suggesting that matching spacers can often be explained by HGT.

While this provides a satisfactory explanation for 52% of observed spacer matches, the remaining 48% remain unexplained. The requirement for 100% identity among repeats for classification as transferred is likely to be excessively stringent. There are many cases of true homology that are excluded by this cut-off. As shown in Figure 2.3, 16% of comparisons are below the 100% identity cutoff but above 80% identity. These cases may also reflect homologous systems that were horizontally transferred, but the evidence is less clear. Meanwhile, the remaining 32% of cases contain truly divergent repeats that cannot easily be interpreted as a consequence of HGT. Instead, these non-homologous CRISPR with identical spacers appear to result from convergent evolution.

Whether matching spacers are acquired through HGT or convergent evolution, both mechanisms suggests positive selection to favor the proliferation of these shared spacers. One simple explanation for such selection would be that these shared spacers target especially common genetic elements that are shared across MGE. To test this hypothesis, we used BLAST to probe a database of 1,633 phage and 1,055 plasmid genomes from GenBank. We identify regions of homology between CRISPR spacers and these MGE genomes (e-value $< 1 \times 10^{-5}$). As shown in Figure 2.4, shared spacers are more than four times as likely to match a sequenced phage and three times as likely to match a sequenced plasmid than unique spacers that are not shared across sequenced isolates ($P < 1 \times 10^{-200}$ in both cases, Chi-Square). Shared spacers target sequences that are especially common among mobile genetic elements. In particular, matching spacers share homology

to phage four times more than plasmids. This is a notable observation, because evidence for long-distance phage interactions has been much more limited than among plasmids (Hyman and Abedon 2010).

Several of the early comparative studies of mobile elements identified nearly identical sequences among the otherwise highly divergent milieu of mobile elements (Frost et al. 2005; Hendrix et al. 1999; Hambly and Suttle 2005). This work indicates that bacteria have discovered these common traits and exploited them using CRISPR spacers that provide broad immunity to common mobile elements. Once identified, these effective spacers have been widely shared via HGT, even among distantly related strains.

Ironically, mobile elements themselves appear to be the engine that drives the distribution of molecular defenses against MGE, as HGT appears to be the most common evolutionary path leading to shared spacers. Indeed, CRISPR carriage may be an effective strategy for MGE to simultaneously exclude competitors and promote their own proliferation by benefitting their hosts.

This culture-independent view of MGE evolution provides further evidence that promiscuous mobile genetic elements interact with a broad range of bacterial hosts. These interactions can be captured by CRISPR arrays, which enable a broad phylogenetic view of bacteria-MGE interactions. This study does not distinguish between the host range of an entire mobile genetic element and individual genes contained within an MGE. However, we believe that as in their bacterial hosts, genes, not genomes are the fundamental unit of selection and evolution among MGE. As a result, we expect that recombination among MGE allows some genes to traverse the bacterial phylogeny independent of their original phage or plasmid host. This provides a mechanism that would select for shared spacers in the absence of directly overlapping MGE host ranges.

Nonetheless, although targeted genes may be passed among many intermediates, the transfers required to achieve the observed phylogenetic distribution still require broad host range MGE. For example, intermediation does not reduce the magnitude of an

exchange between gram negative and gram-positive hosts. Matching CRISPR spacers in distantly related hosts provide additional evidence of broad host-range MGE, countering the traditional, culture-based view of MGE host range.

2.3 Methods

Data Sources: All 9,472 bacterial genomes analyzed in this work were accessed through GenBank (Benson et al. 2005). The 1,633 phage and 1,055 plasmid genomes used in this work were downloaded from EBI (Flicek et al. 2011).

CRISPR Identification: The CRISPR Recognition Tool (CRT) was used with default parameters to identify CRISPR arrays in all of the available bacterial genomes (Bland et al. 2007).

16S rRNA Distance Matrix: We first filter out genomes that lack a full length (>1000 bp) hit to the greengenes database (DeSantis et al. 2006a) and genomes that do not have a CRISPR array, leaving 3,314 genomes in our analysis. For each of these genomes, we use PyNAST (J. G. Caporaso, Bittinger, et al. 2010) to create a profile alignment, followed by a global alignment using the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) to determine the pairwise distance between profile alignments of each 16S rRNA sequence.

Spacer Comparisons: Synthetic spacers were generated using custom python scripts and comparisons among spacers and repeats were conducting using BLAST with low-complexity filtering disabled. Significant spacer matches were determined by hits with 100% identity over at least a 30 base-pair window. Significant homology to phage and plasmid genomes was determined by filtering for hits with an e-value below 1×10^{-5}.

Repeat Analysis: For comparisons of CRISPR repeats, when multiple spacers are shared across a pair of genomes, the pair with the higher percent identity is used, to provide the most sensitive detection of transferred CRISPR. For comparisons that do not share a matching spacer, a consensus repeat sequence is used.

2.4 Figures

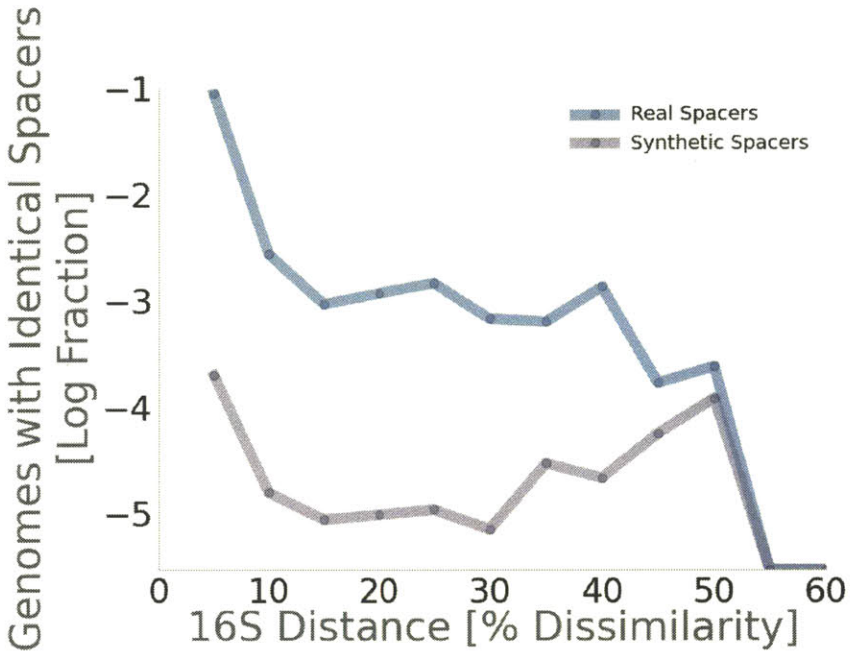


Figure 2.1: Distantly related bacteria share identical spacers. The fraction of genome comparisons containing identical spacers (blue) or synthetic spacers (purple) is shown across all phylogenetic distances. To facilitate comparison, fractions are plotted on a log scale, with zero values set to -5.5 to enable comparison with other values. Values are binned across intervals of 5% 16S rRNA distance to reduce noise. Genuine spacers are shared more frequently than synthetic spacers.

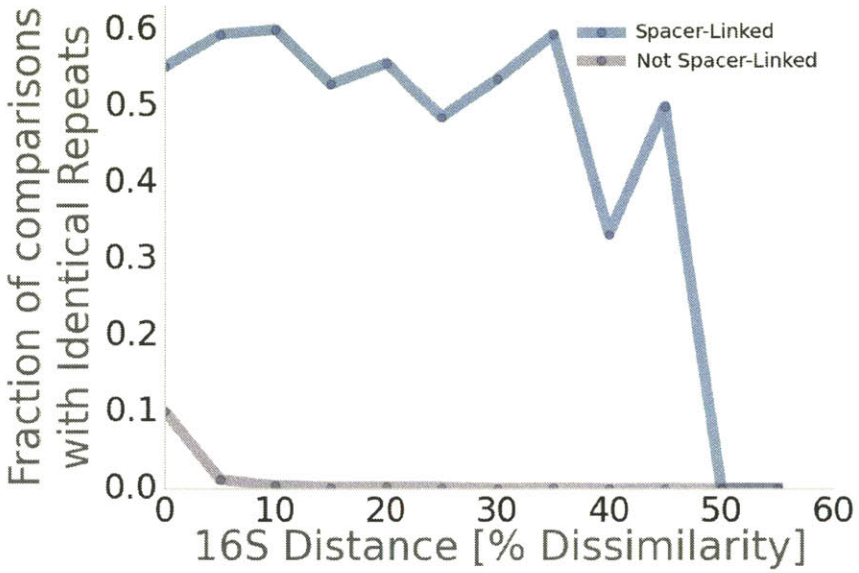


Figure 2.2: Shared spacers frequently have identical repeats. Among all genomes that either share a matching spacer (blue) or do not (purple), the fraction of comparisons that have an identical repeat is presented. Comparisons are discretized into 5% distance bins. Genomes with matching spacers are much more likely to have matching repeats, suggesting that in many cases, matching spacers may be caused by horizontal gene transfer.

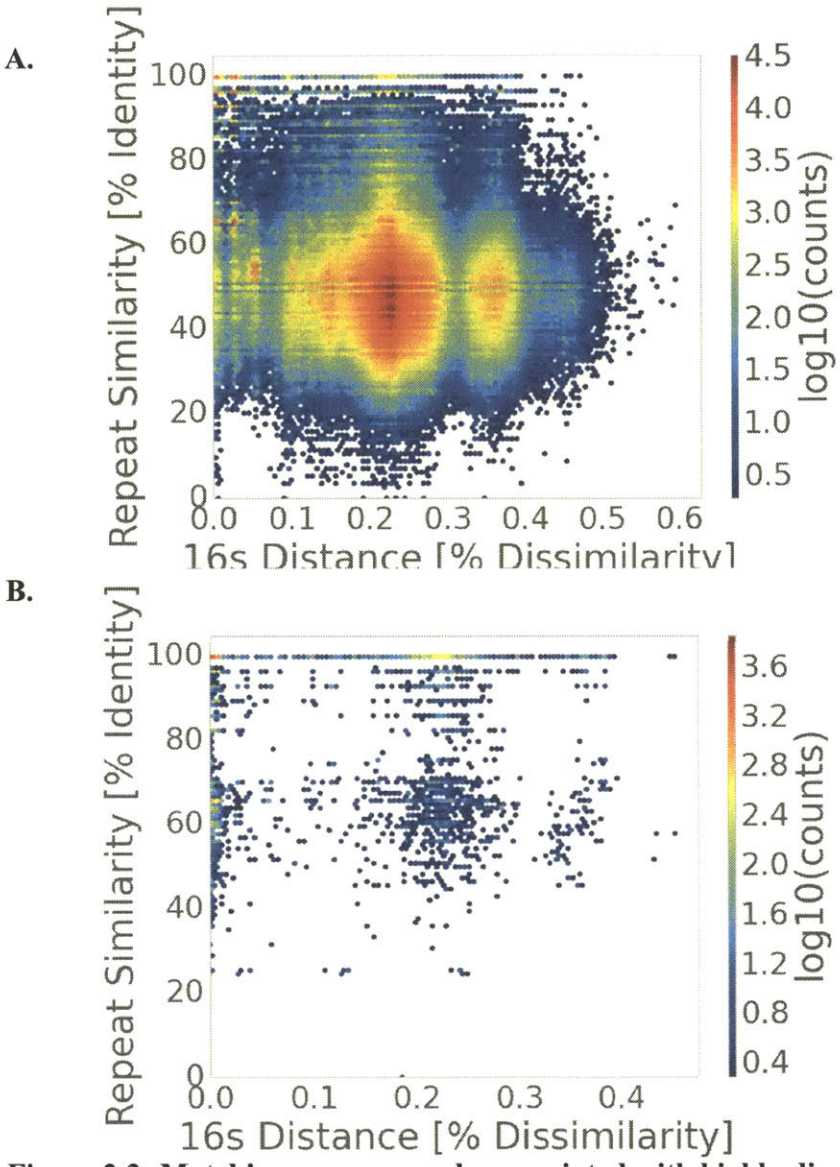


Figure 2.3: Matching spacers can be associated with highly divergent repeats. The identity of repeats is shown across 16S rRNA distance for pairs of genomes without matching spacers (**A**) and for genomes with matching spacers (**B**). The number of counts in each bin is colored on a log scale according to the colorbar in each panel. Genomes without matching spacers typically have repeats with low identity. A much higher fraction of genomes with a matching spacer have identical repeats, however there is a significant sub-population with highly divergent repeats that is unlikely to be homologous. This sub-population of matching spacers with repeats below 80% identity are not likely to be explained by HGT.

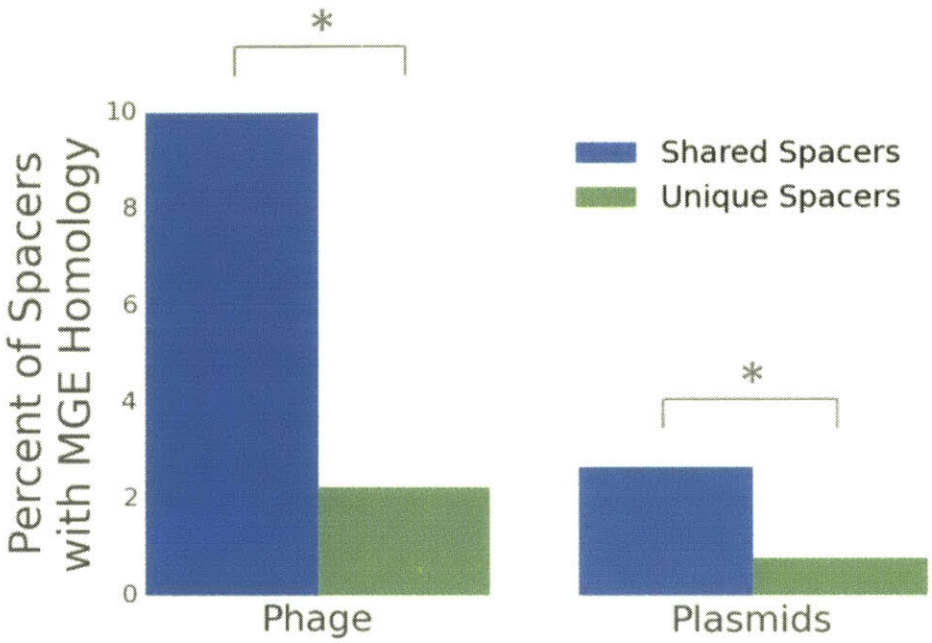


Figure 2.4: Shared spacers matched sequenced MGE more than unique spacers. The percent of spacers with homology to a sequenced phage or plasmid are shown for shared spacers with an identical match in another genome (blue) and unique spacers (green). Statistically significant comparisons are marked with an asterisk ($P < 1 \times 10^{-200}$, Chi-Square). Shared spacers are more likely to share homology to a sequenced MGE, suggesting that these elements target especially common components of mobile elements.

Chapter 3: Natural bacterial communities as quantitative biosensors

Smith MB, Rocha, AM, Oleson, SW, Paradis CJ, Smillie, CS, Campbell, JH, Forney, JL, Mehlhorn, TL, Lowe, KA, Earles, JE, Phillips, J, Techtmann, SM, Joyner, DC, Preheim, SP, Sanders, MS, Mueller, MA, Brooks, S, Watson, DB, Wu, L, Zhang, P, He, Z, Zhou, J, Adams, MW, Lancaster, A, Poole, F, Adams, PD, Arkin, AP, Fields, M, Alm, EJ, Hazen, TC, (2014) Natural bacterial communities as quantitative biosensors (*In progress*)

3.1 Abstract

Human impacts on the environment ranging from the production of persistent nuclear waste to more transient oil spills threaten ecosystem stability and in turn, human welfare. Uncovering contamination facilitates remedial action. Here we show that analysis of DNA from natural bacterial communities can be used to accurately identify environmental contaminants including uranium and nitrate at a nuclear waste site. We show that beyond contamination, 16S rRNA sequence data alone can quantitatively predict a rich catalogue of 26 geochemical features collected from 93 wells with highly variable geochemistry. We extend this approach to identify sites contaminated with hydrocarbons from the Deepwater Horizon oil spill. These results indicate that bacterial communities can be used as environmental sensors that respond to, and capture perturbations caused by human impacts.

3.2 Main Text

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 (Moss et al. 2010). 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. 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 (Fischer, Agarwal, and Hess 2009; Wu et al. 2011) or even whole bacterial cells (Belkin 2003) can be used as biosensors to translate environmental signals into machine-readable data (Su et al. 2011; D'Souza 2001). However, these systems must be carefully engineered before use and cannot be deployed in environments that are unsuitable to the particular proteins or cell lines being utilized. Rather than using a single macromolecule or strain, here 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 will predictably restrict or promote the growth of characteristic taxa in accordance with environmental conditions, a basic hypothesis that is central to ecological theory (Darwin 1859). Consistent with this model, previous efforts have uncovered correlations between the composition of bacterial communities and environmental features such as pH (Lauber et al. 2009) or temperature (Gianoulis et al. 2009). These and many other descriptive efforts are based on correlations fit directly to observed data rather than cross-validated models suitable for predictive use, leaving indigenous biosensors largely unexplored.

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. As a result of the unusual chemical processes that were performed here, this site harbors spectacular geochemical gradients. For example, at some locations, pH varies by 7 units over 30 meters.

We endeavor to use modern machine learning tools to translate 16S rRNA sequence data from native bacterial communities into a predictive model that can discriminate between wells that have concentrations of uranium or nitrate - the two primary pollutants in this watershed - that are either above US standards for safe drinking water (contaminated) or below the standard (not contaminated). We chose to build these contamination models with Random forest (Liaw and Wiener 2002) after evaluating nine machine learning tools (Pedregosa et al. 2011) trained and tested on data generated from this site (see Supplemental Figs. 3.2-3.5). Random forest is an ensemble-based supervised machine learning algorithm that has already been successfully used to classify disease and other host-microbe relationships from 16S rRNA sequence data (Papa et al. 2012; Metcalf et al. 2013).

To inform this model, we collected extensive geochemical data and DNA for sequencing. 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-medians clustering We physically sampled one site from each of the 100 resulting clusters, excluding 7 clusters that were inaccessible (see Supplemental Fig. 3.1 and Supplemental Table 3.1). Groundwater from each site was accessed at a mean depth of 11.4 meters using monitoring wells drilled throughout the watershed. At each of these selected sites we measured 38 geochemical and physical features (see Tables S2-S3).

Bacterial communities within each well were collected on a 10 μm filter to retain particle attached cells and a 0.2 μm 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 using a distribution-based clustering algorithm (Preheim et al. 2013) to reduce redundancy, we still observe 26,943 unique operational taxonomic groups, 9,306 of which have not been previously characterized, highlighting the unusual biological diversity of this site. Prior to prediction of contamination, we

filtered low-abundance and narrowly distributed taxa, yielding 2,972 operational taxonomic units as features.

We find that our contamination classifier - trained on 16S rRNA data alone - is able to accurately distinguish between safe sites and those contaminated with either uranium (F1 score = 0.88) or nitrate (F1 = 0.73). Fig. 3.1 shows the distribution of wells sampled across the contaminant gradients as well as classifier performance at each site. Despite nearly equal representation of features from both free-living (0.2 μm filter, 1554 taxa) and particle attached (10 μm filter, 1418 taxa) communities, the preponderance (88%) of features important for predicting uranium are from the free-living fraction ($p < 10^{-6}$, Fisher's exact test). No such enrichment is observed for nitrate classification, suggesting that the effect is specific to the biology of uranium-responsive taxa. As further evidence of ecological stratification across size fractions, we are able to accurately predict which size fraction otherwise identical samples are drawn from using our supervised machine learning approach (see Supplemental Fig. 3.6). 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 is not significantly impacted (see Supplemental Figs. 3.7-3.9).

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, *Methylococcus* and *Brevundimonas* are among the most informative features for nitrate classification and both are known to be active nitrate reducers (Kavitha et al. 2009). Similarly, the most

important features for identifying uranium contamination included *Rhodanobacter* and *Rhodocyclaceae*, both of which have been previously identified for their role in uranium reduction and bioremediation (Green et al. 2012). These results suggest direct ecological associations can be used to accurately identify environmental contaminants.

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 lack direct ecological associations (e.g., aluminum) but that are instead correlated with other forces that *are* ecologically relevant (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 directly informs 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, here we predict 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 data can accurately predict pH, recovering spatial variance across the site (see Fig. 3.2), with a significant correlation between predicted and true values ($p < 10^{-10}$, $\tau = 0.46$, Kendall tau rank correlation). Out of 38 total 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 is not believed to play an important role in biological systems ($p < 10^{-10}$ and $p < 0.005$ respectively, Kendall tau rank correlation, Fig 3.2). Although biologically relevant traits like pH can be predicted more accurately than traits with less direct ecological impacts, we find that natural bacterial communities create a broadly informative imprint of their environment.

To explore whether this approach can be applied in other ecosystems and perturbations, we analyzed previously reported data (Hazen et al. 2010) collected before and after the 2010 Deepwater Horizon oil spill in the Gulf of Mexico. In the worst marine oil spill in US history, 4.1 million barrels of crude oil were released 1500 meters below the surface, 80 kilometers 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 of 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. 3.3). As a demonstration of the general utility of this approach, these data were collected with an unrelated DNA measurement technology, using a PhyloChip 16S rRNA microarray, rather than through 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 either our uranium or nitrate classifiers (see Fig. 3.3).

To explore the ecological mechanisms that may underlie this surprisingly effective oil biosensor, we consider the niches of two particularly well-studied predictive features. Oceanospirillaceae is a clade containing many known hydrocarbon degrading specialists (Hazen et al. 2010; Teramoto et al. 2011) that is highly enriched in oil-contaminated sites. Pelagibacteraceae is an oligotrophic clade that dominates nutrient-poor aquatic environments, is thought to be among the most abundant organisms on earth, and is enriched in uncontaminated sites in our dataset (Morris et al. 2002; Carini et al. 2013). Consistent with these distinct niches, the oil biosensor is informed both by an enrichment of Oceanospirillaceae and a depletion of Pelagibacteraceae among contaminated sites. The relative abundance of these two organisms alone is sufficient to accurately discriminate between contaminated and uncontaminated sites (Fig. 3.4b).

Interestingly, in this plot of the oligotrophic Pelagibacteraceae and the oil-degrading Oceanospirillaceae, 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 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 equally as well as truly contaminated sites (F1 = 0.98) even though the oil itself is missing at these locations. This indicates that the ecological signatures of human interventions can persist beyond the depletion of geochemical markers.

To determine the phylogenetic breadth needed to build an effective indigenous biosensor, we compare the phylogenetic distribution of the most predictive features for oil and uranium, our two best classifiers (Fig. 3.4). There are significant phylogenetic associations between these features and the data they predict. Betaproteobacteria are enriched among uranium-predictive features ($p < 0.01$, fisher exact test), and Gammaproteobacteria are enriched among oil-predictive features ($p < 0.001$). However, beyond these groups, there is considerable variation, with highly predictive features for both oil and uranium interleaved throughout the rest of the tree of life, highlighting the phylogenetic diversity of taxa associated with these contaminants.

Much previous work has focused on using bacteria to report useful information from their environment. However, these efforts have typically used electrochemical or optical properties of well-characterized strains in response to defined targets that are typically metabolized. In contrast, this study shows that with appropriate training data and analytical models, natural bacterial communities can be used as biosensors for a broad array of geochemical measurements, including many that 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.

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 with the constraints of existing sequencing technology, indigenous biosensors may already be well suited for applications where it is possible to fully

characterize a small training set, and necessary to loosely monitor a broad range of geochemical features over a very large sample size. 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 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 to most organisms because of its complex chemical structure. It is a rich reward for specialists like 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 only becomes ecologically relevant in the presence of a suitable carbon source, which is rare in highly oligotrophic groundwater communities. As a result, the selective advantage is less significant than for 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 (Papa et al. 2012).

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 less than 93 labeled samples. Given ubiquitous, ecologically structured gene exchange (Smillie et al. 2011), we expect that many ecological associations will be captured in the flexible gene pool. Consequently, a richer set of features comprised of shotgun metagenomics or single-cell genomes should yield more powerful classifications. Transcriptomic data could capture instantaneous responses to environmental 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. Finally, these results highlight the broad, lasting nature of human impacts on the environment - bacterial communities today continue to faithfully report the impact of nuclear waste created at the dawn of the atomic age.

3.3 Methods

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 for comparison located within the Bear Creek Valley watershed in Oak Ridge, Tennessee. Contamination at this site includes radionuclides (e.g. Uranium, Technetium), nitrate, sulfide, and volatile organic compounds (Watson and Kostka 2004). The main source of contamination is 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 (Watson and Kostka 2004). These source plumes are continuously monitored and have been the subject of a number of studies over the years (Green et al. 2010; Green et al. 2012). Further information regarding the plume and sources of contamination can be found at http://www.esd.ornl.gov/orifrc/.

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 result of 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.

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. Supplemental Fig. 3.1 illustrates the high diversity of 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, $p < 1e\text{-}10$, 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.

Geochemical and physical measurements

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 non-contaminated background wells, with each well representing a distinct geochemical transect.

All groundwater and filtered-groundwater samples were collected from mid-screen 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 2-20L of groundwater was purged from each well. For all wells, water was collected with either a peristaltic or bladder pump using low-flow in order to minimize drawdown in the well.

A total of 15 geochemical and 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 (In-situ Inc., Colorado). To ensure accuracy, dissolved oxygen and pH probes were calibrated daily and the remaining probes calibrated monthly. Sulfide and Ferrous Iron (Fe(II)) groundwater concentrations were determined using the USEPA Methylene Blue Method (Hach 8131) and 1,10-Phenanthroline Method (Hach 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 (APHA 2012), unless otherwise indicated. A description of the sampling and analytical methods for each parameter is provided in the following sections.

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, N_2O) were measured on a SRI 8610C Gas Chromatograph with Argon carrier gas, using a method derived from EPA RSK-175 and 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 were collected in pre-cleaned volatile organic analysis (VOA) vials with no headspace, and stored upside down at 4°C until analyzed. To minimize diffusion of oxygen into the VOA vials through the septa, samples were analyzed within 5 days. The day of analysis, samples were brought up to room temperature and weighed (vial + cap + 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 minutes and

vials re-weighed with the headspace. Gas samples were withdrawn using a gas tight syringe (within 3 minutes after shaking has stopped). The sample was injected into a gas chromatograph for analysis and peak areas compared to known standards to calculate the quantity of each gas.

Dissolved Carbon

Dissolved Organic Carbon (DOC) and 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 40mL pre-cleaned VOA vials with no headspace. To determine DIC, the samples were placed on the autosampler and inorganic carbon was measured as CO_2 is released in the TOC analyzer. To determine concentrations of DOC, the samples were acidified with 2N HCl and sparged with high-purity oxygen to remove the inorganic carbon. Samples were then injected onto the combustion chamber of the carbon analyzer and the resulting CO_2 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 2-200 ppm and .5-100 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 one week of collection. All reagents were prepared following EPA method protocols.

Anions

Anions (bromide, Chloride, nitrate, phosphate, and sulfate) were determined using a Dionex 2100 with an AS9 column and carbonate eluent (Method # here). The Dionex uses chromatographic separation and conductivity to measure concentration compared with a standard curve. To determine anions, 20-mL of filtered groundwater (0.22 μm filter unit) was collected in 20mL plastic scintillation vials with no to little headspace and stored at 4°C until analyzed. For analysis, the sample was loaded and 10μl injected into the instrument column. Calibration curves for each analyte were prepared using standard concentrations.

Metals

Detection of metals (and trace elements) in the groundwater were determined on an Inductively Coupled Plasma/Mass Spectrometry (ELAN 6100 ICP-MS) using a method similar to the EPA method 200.7. For determination of dissolved elements, filtered groundwater samples (0.22 μm filter unit) were collected in certified sterile VWR® Metal-Free (<1ppb for critical trace metals) polypropylene centrifuge tubes and stored on blue ice until transported back to the laboratory. At the lab, the 0.1 – mL of each sample aliquoted into a new VWR Metal-free tube and diluted with 1% nitric acid solution to preserve the sample (pH <2) A multi-elemental internal standard is added directly to the diluted sample. A set of multi-element calibration standards is prepared to cover the desired range of analysis. Next, samples are introduced into the system using a peristaltic pump and PerkinElmer model AS-93 auto-sampler. To ensure quality control, a duplicate and matrix spike samples were included in every run (approximately 1 per every 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/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 analyzed. A Corning MP-3A distillation apparatus was used to produce pure glass distilled water (gddH_2O) used in all dilution and washing steps. Tubes used in ICP-MS analysis were acid-washed by submersion in an 2% v/v solution of concentrated nitric acid in gddH_2O for 24 hours and rinsed twice by submersion in pure gddH_20 for 24 hours. Trace metal grade concentrated (70%) nitric acid (Fisher A509-212) was used in acidification of samples. To measure both soluble and insoluble elements present, groundwater samples (6ml) were placed into acid-washed 17x20mm Sarstedt polypropylene screw-cap conical tubes (62.554.002-PP) and centrifuged at 7,000xg for 15 minutes 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% v/v) of concentrated nitric acid. To the pellet was added 6 ml

of 2% v/v concentrated nitric acid in gddH_2O. All samples were briefly vortexed (30 seconds) and incubated at 37°C for 1 hour in a New Brunswick Scientific G24 Environmental Incubator Shaker with shaking speed setting of medium. All samples were centrifuged at 2000xg for 10 minutes in a Beckman Allegra 6R centrifuge at 25°C. Metal analysis of all samples was performed in triplicate using an Agilent 7500ce octopole ICP-MS in FullQuant mode using and internal standard with in-line addition and multi-element external standard curve as previously described (1). Samples were loaded via a Cetac ASX-520 autosampler. Control of sample introduction, data acquisition and processing was performed using Agilent MassHunter version B.01.01.

Direct Cell Counts

Bacterial biomass in groundwater samples was determined using the acridine orange direct count (AODC) method (Hazen et al. 2010). For each well, 40-mL of groundwater samples were preserved in 4% formaldehyde (final concentration) and stored at 4°C. To prepare slides, 1-10mLs of groundwater were filtered through a 0.2-μm black polycarbonate membrane (Whatman International Ltd., Piscataway, NJ). Filtered cells were then stained with 25mg/ml of Acridine orange (AO), incubated for 2 minutes in the dark, and filtered again to remove any unbound acridine orange stain. The filters were rinsed with 10-mL of filter-sterilized 1XPBS (Sigma Aldrich Corp., St. Louis, MI) and the rinsed membrane mounted on a slide for microscopy. Cells were imaged using a FITC filter on a Zeiss Axio Scope A1 (Carl Zeiss, Inc., Germany).

DNA collection and extraction

DNA was collected by sequentially filtering 4-L of groundwater through a 10.0-μm Nylon pre-filter and 0.2μm- Polyethersulfone (PES) membrane filter (144mm diameter, Sterlitech Corporation). Filters were stored in 50-mL falcon tubes and immediately stored on dry ice until transported back to the laboratory. At the laboratory, samples were stored at -80 °C until extracted using a modified Miller method (J. Caporaso et al. 1999; Hazen et al. 2010). For each sample, the filter was cut in half and each half placed into a Lysing Marix E tube (reduced to 50% of the tubes; MP Biomedicals, Solon, OH). 1.5mL of Miller phosphate buffer and Miller SDS lysis buffer were added to each tube and mixed.

Next, 3.0mL of phenol: chloroform: isoamyl alcohol (25:24:1) and 3.0mL of chloroform were added to each tube. The tubes were bead-beat at med-high speed for 5 minutes. The entire contents of the tube were transferred to a clean 15-mL Falcon tube and then spun at 10,000× g for 10 minutes at 4°C. The upper phase (supernatant) was transferred to a clean 15-mL tube an equal volume of chloroform was added. Tubes were mixed and then spun at 10,000 ×g for 10 min, aqueous phase (~2-3mL) was transferred to another tube and 2 volumes of Solution S3 (MoBio Power Soil, Carlsbad, CA) was added and mixed by inversion. 650 µl of sample was loaded onto a spin column and filtered using a multi-filter vacuum apparatus. This was continued until all the solution was filtered. Next, 500µl of Solution S4 (MoBio Power Soil, Carlsbad, CA) were added to each filter then spun down at 10,000×g for 30 seconds. The flow-through was discarded and spun for another 30 seconds to ensure the all solutions had been filtered. Samples were recovered in 100µL Solution S5 (MoBio Power Soil, Carlsbad, CA) and stored at -20°C.

Library preparation and sequencing

Polymerase Chain Reaction (PCR) Primers

A two-steps PCR amplification method was used for PCR product library preparation to avoid extra PCR bias to be introduced by Illumina adapter and other added components. Standard primers [515F, 5'-GTGCCAGCMGCCGCGGTAA-3' and 806R, 5'-GGACTACHVGGGTWTCTAAT-3' targeting the V4 region of both bacterial and archaeal 16S rRNA without added components were used in the first step PCR.

To increase the base diversity in sequences of sample libraries within V4 region, phasing primers were designed and used in the second step of the two-step PCR. Spacers of different length (0-7 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 length of the amplified sequences do not vary with the primer set used, the forward and reverse primers were used in a complementary fashion so that all of the extended primer sets have exactly 7 extra bases as the spacer for sequencing phase shift. Barcodes 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 barcodes (12 bases), the Illumina reverse read sequencing primer (35 bases),

spacers (0-7 bases), and the target reverse primer 806R (20 bases). The forward phasing primers included (from 5' to 3') an Illumina adapter for forward PCR (25 bases), the Illumina forward read sequencing primer (33 bases), spacers (0-7 bases), and the target forward primer 515F (19 bases).

PCR amplification and purification

In the first step PCR, reactions were carried out in a 50 μl reaction: 5 μl 10×PCR buffer II (including dNTPs), 0.5 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse target only primers, 10 ng soil DNA or 1 μl mock community of 20x dilution. Samples were amplified 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, USA), eluted in 50 μl water, and aliquoted into three new PCR tubes (15 ul each). The second round PCR used a 25 μl reaction (2.5 μl 10×PCR buffer II (including dNTPs), 0.25 U high fidelity AccuPrime™ Taq DNA polymerase (Life Technologies), 0.4 μM of both forward and reverse phasing primers, 15 μl aliquot of the first-round purified PCR product). The amplifications were cycled 20 times following the above program. 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 *molality*. The pooled mixture was purified with a QIAquick Gel Extraction Kit (QIAGEN Sciences, Germantown, MD, USA) and re-quantified with PicoGreen.

Sequencing

Sample libraries for sequencing were prepared according to the MiSeq™ Reagent Kit Preparation Guide (Illumina, San Diego, CA, USA) as described previously (J. G.

Caporaso et al. 2012). 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. 980 µL of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, the 20pM 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 ends reads. The 16S rRNA library for sequencing was mixed with a about 10% Phix library (final concentration).

A 500-cycle v1 or v2 MiSeq reagent cartridge (Illumina) was thawed for 1 h in a water bath, inverted ten 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.

Data processing

Initial filtering and processing

16S rRNA sequence data generated from MiSeq were processed to overlap paired-end reads and to filter out poorly overlapped and poor quality sequences. Sequences were de-multiplexed 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 was divided using a custom script (split_fastq_qiime_1.8pl) to facilitate parallel processing with SheRA (http://almlab.mit.edu/shera.html) (Rodrigue et al. 2010). 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). Fastq format was regenerated from the resulting fastq and quality files with mothur (version v.1.25.0) make.fastq command (default parameters, including sanger ascii offset 33 scaling) (Schloss et al. 2009). 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 phred/ascii offset of 33 (--phred_offset 33) (J. G. Caporaso, Kuczynski, et al. 2010). Finally, the primer sequences and any sequence outside of the amplified region was removed using a custom script (remove_primers_staggered.pl).

Creating operational taxonomic units

Operational taxonomic units (OTUs) were generated as previously described with either distribution-based clustering (DBC) or USEARCH (usearch_i86linux32 v6.0.307, drive5.com) (Preheim et al. 2013; Robert C. Edgar 2010). First, the sequences were truncated at 251 bp (truncate_fasta2.pl), de-replicating 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 dataset. For DBC, de-replicated (100% clusters), filtered sequences were progressively clustered with UCLUST (drive5.com) to 94% identity. DBC was run as previously described (Preheim et al. 2013) from the 94% identity pre-clustered 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). Representative sequences for each OTU are the most abundant sequence within the OTU.

Classification and removal of chimeras and non-specific sequences

OTU with representative sequences that are chimera or a non-specific amplification product 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 (drive5.com) with default parameters

and removed. Finally, sequences were classified using the Ribosomal Database Project classifier (version 2.3) (Cole et al. 2005).

Identification of novel OTUs

To determine the fraction of OTUs at this site that have not been previously characterized, we BLASTed a representative sequence from each OTU against the most recent release of the GreenGenes 16S rRNA database (Version 13_5) using 97% identity gene clusters (DeSantis et al. 2006b). For each OTU, we considered the highest identity nearly full-length hit (>250 bp). If the best hit to the GreenGenes database was at least 95% identical, we considered the OTU previously characterized. 9,306 of the 26,943 OTUs in our dataset did not have a hit in the GreenGenes database with at least that identity. We consider these OTUs to be novel. We chose the 95% cut-off as a conservative alternative to the typical 97% cut-off used for identifying OTUs. A 97% cut-off yielded 13,371 novel OTUs. A 93% cut-off yields 5,925 novel OTUs.

Machine Learning

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 simple 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 rRNA data from those wells. We chose to use the popular scikit-learn machine learning toolkit (Pedregosa et al. 2011) to run the experiment, as this enabled us to quickly swap between 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 over-all performance in terms of training time, cross-validated accuracy. We also considered that Random forest has been widely used in the literature (Papa et al. 2012; Metcalf et al. 2013) 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, out 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 non-ensemble decision tree method proved to be better than the linear models on average, 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.

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 filter dataset and 1,419 OTUs from the 10 μm filter dataset. We concatenate these matrices to allow information about the niche-specific abundance of each of these OTUs to inform our model. Prior to

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 this supplement. All reported results are trained with 1000 trees. Reported accuracies reflect the out-of-bag error for each run of the random forest. Performance metrics are computed from a confusion matrix populated by out-of-bag predictions. ROC 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.

Permutation testing

To validate our machine learning pipeline, we subjected a sub-sample of predictions to a permutation test. Specifically, we have randomized the labels associated with real training data and performed all down-stream analysis as usual to determine whether our predictions could be explained by chance due to some inherent structure to the data. We performed this control to observe the variance in predictions achieved by shuffled data

and with 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 rRNA data-set (both data from the 0.2 μm and 10 μm 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 AUC (area under the receiver operator curve) 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 Supplemental Fig. 3.5.

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, however on average, the random association between labels and the training data does not allow accurate classification.

Evaluating geographic confounds

Geographic Structure at Oak Ridge

As illustrated in Figure 3.1 in the main text, there is considerable geographic structure to 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 over-fitting 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.

Data-filtering

Geographic over-fitting is most likely among taxa that are geographically restricted. As one methodological control against this type of over-fitting, we pre-filter 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.

Evaluating the relationship between feature proximity and geographic distance

As a first step towards evaluating the effect of geography on contaminant classification, we have computed the feature-space proximity for all wells using the random forest package (Liaw and Wiener 2002). 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 feature space. In Supplemental Fig. 3.7, 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 Supplemental Fig. 3.8.

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.

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-neighbor model as a null against which to compare these results. A general assumption of supervised machine learning tools like random forest, is that the training examples are independent from the test sets that are being evaluated. This assumption is violated to some degree in any environmental sampling effort and it is difficult to know *a priori* at what spatial scale samples might stop behaving independently.

Here we endeavor 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 meters. Performance decreases as this radius increases (see Supplemental Fig. 3.9). However, it is difficult to interpret the significance of this observation without paired observation of a null model based purely on geographical proximity.

A nearest-neighbor null-model for evaluating geographic sensitivity

Towards this end, we created a simple nearest-neighbor 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). As expected given the significant geographic structure of this site, this nearest neighbor model performs well, correctly predicting 86% and 77% of well labels for the nitrate and Uranium contamination problems respectively. However, this nearest-neighbor model is highly sensitive to the same geographic exclusion procedure applied to our random forest model above (see Supplemental Fig. 3.9). This suggests that although

the random forest model is sensitive to geographic exclusion, the effect is much smaller than expected from a geography-only model.

Conclusion

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.

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 was 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) were provided by the United States Geological Survey (USGS) National Hydrography Dataset (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.

3.4 Figures

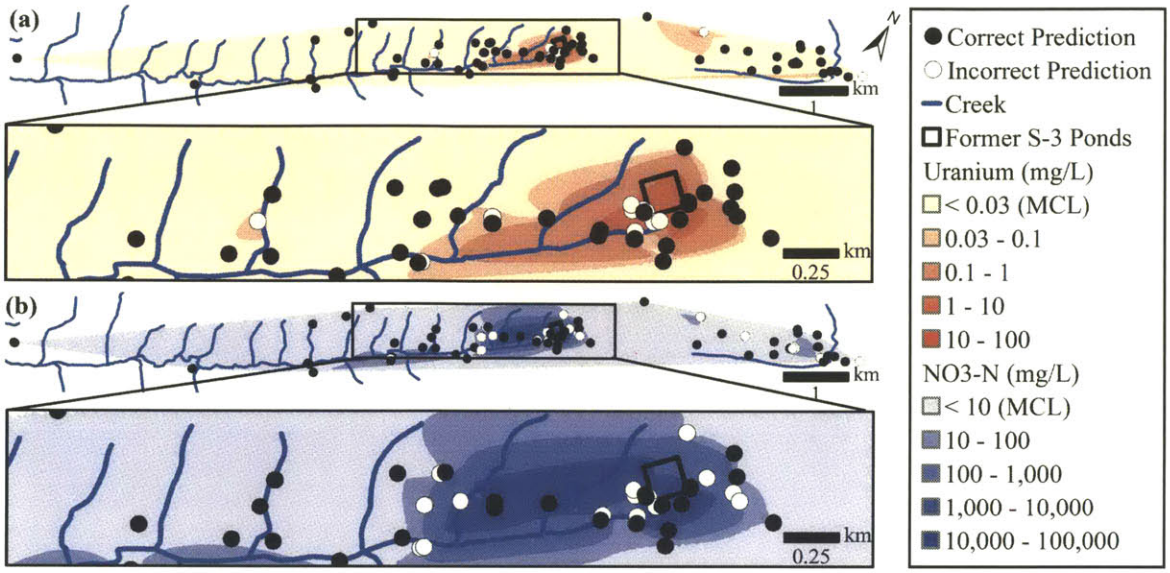


Fig. 3.1: Uranium and nitrate contamination can be effectively identified using bacterial DNA. We trained a Random forest classifier using 16S rRNA abundance data from 2972 operational taxonomic units measured across 93 wells. Classifier performance for uranium (**a**) and nitrate (**b**) is shown across the Oak Ridge Field Site. The Maximum Contaminant Level (MCL) is the cut-off used to determine which sites are contaminated (samples below the cut-off are uncontaminated). Contaminant levels are measured at each well and linearly interpolated between wells. Overall classification performance measured by specificity, sensitivity and accuracy were higher for detecting uranium contamination (0.71,0.87 and 0.82 respectively) than for nitrate (0.81, 0.63 and 0.70).

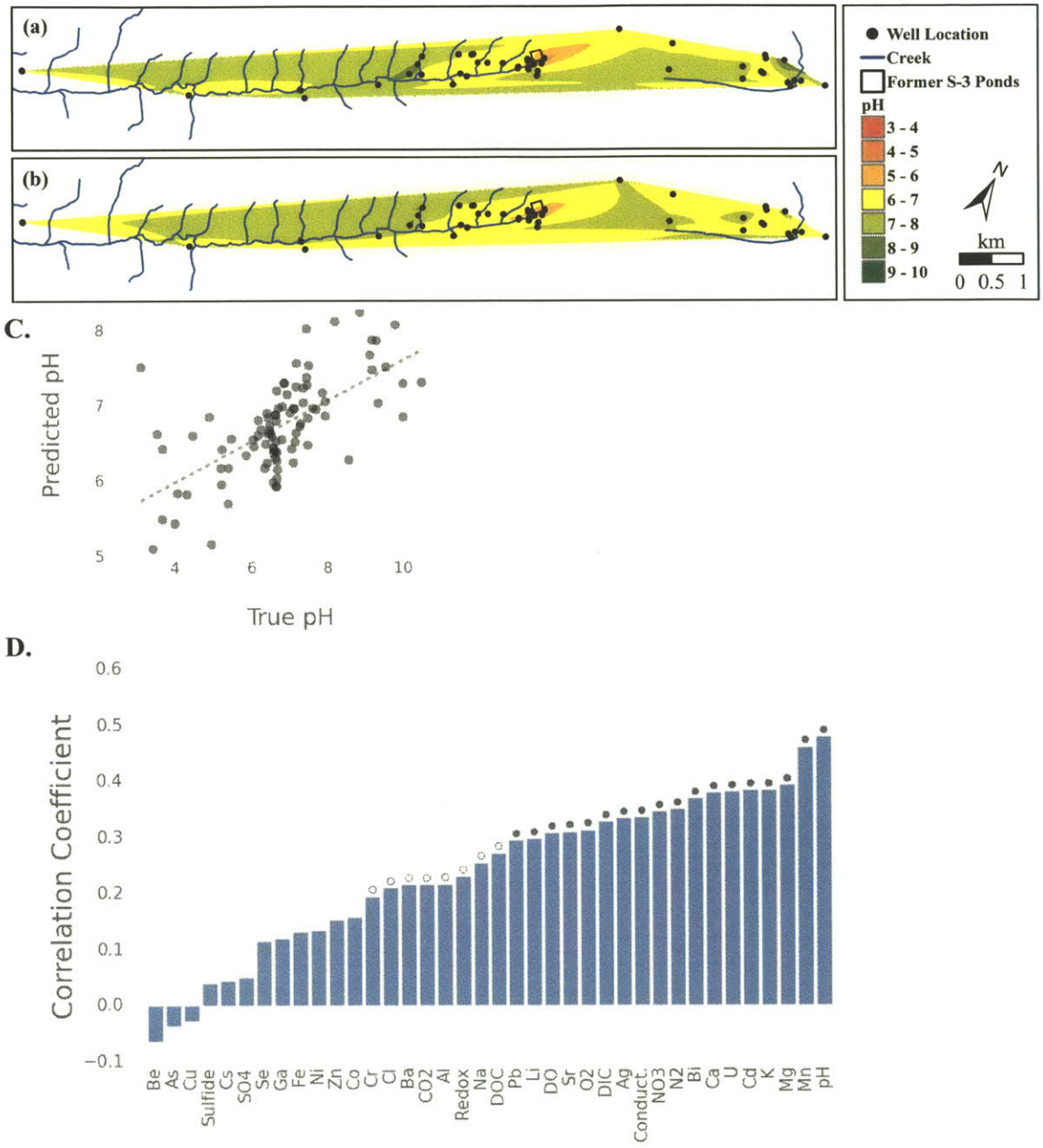


Fig. 3.2: Bacterial DNA can be used to quantitatively predict many geochemical features. Besides classification, we can use 16S rRNA sequence data to predict quantitative values for a variety of geochemical measurements at each well. For example, the prominent features displayed in our map of true pH (**a**) are recovered in our map of predicted pH (**b**). We find that predicted values for pH (**c**) are highly correlated with true values ($p < 1 \times 10^{-10}$, kendall tau rank test). We extended this approach to 38 other geochemical parameters (**d**), where we have plotted the correlation coefficient (Kendall's tau) between true and predicted values. 18 of these correlations are highly significant ($p < 0.0001$, indicated by ●), 8 are significant ($p < 0.01$, indicated by ○) and 12 of these correlations are not significant.

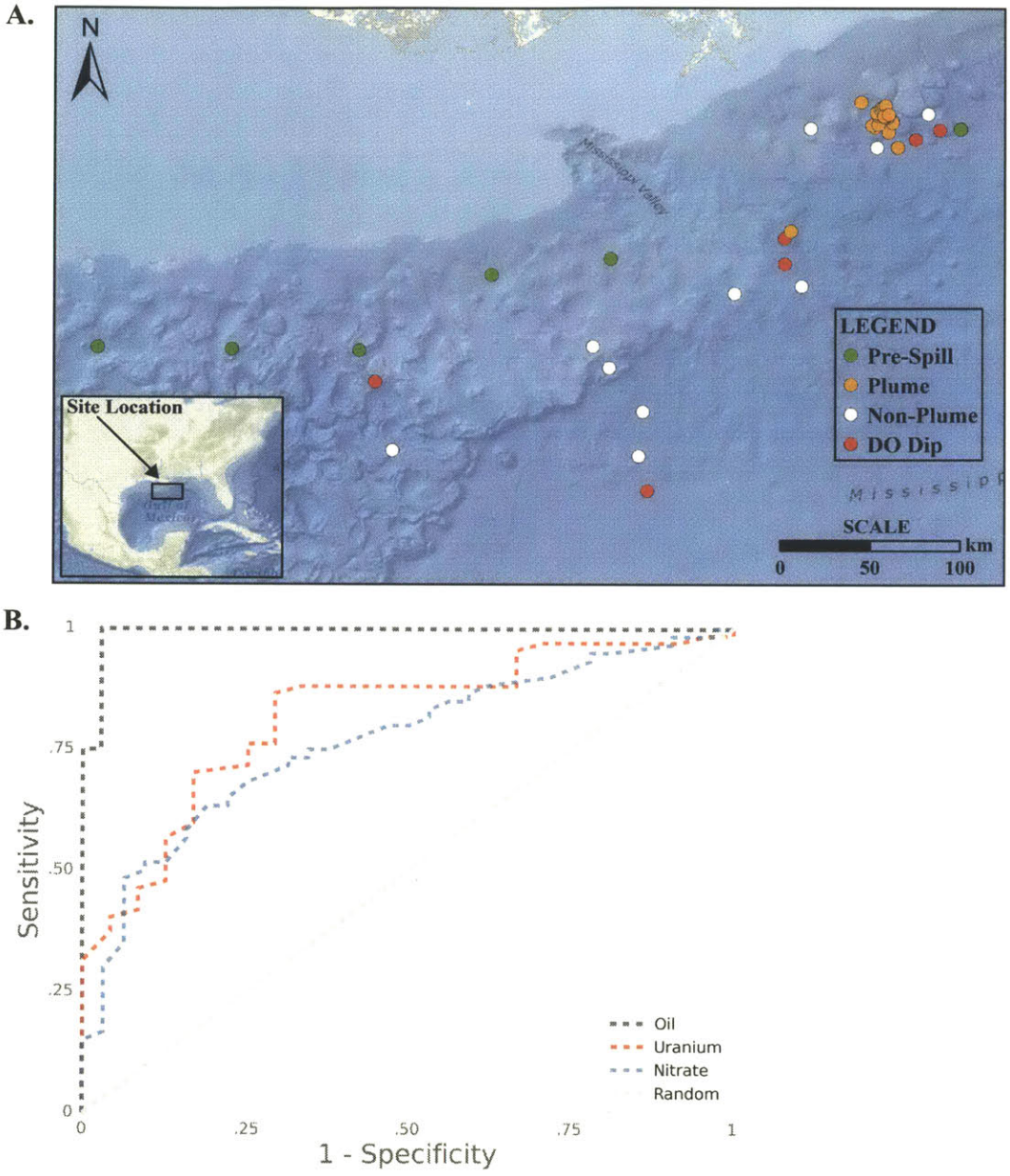


Fig. 3.3: Near-perfect classification of oil contamination using bacterial DNA. Samples collected prior to the Deepwater Horizon oil spill (green), during the spill, but outside of the oil plume (white), from the oil plume (orange) and from the plume but after the oil had been degraded (red) are shown across the Gulf of Mexico (**a**). To compare oil classification performance with classification of uranium and nitrate, we show the receiver operator curves for all classifiers (**b**). The area under the curve is 0.99 for oil, 0.82 for uranium and 0.76 for nitrate, compared to 0.50 for an uninformative random classifier.

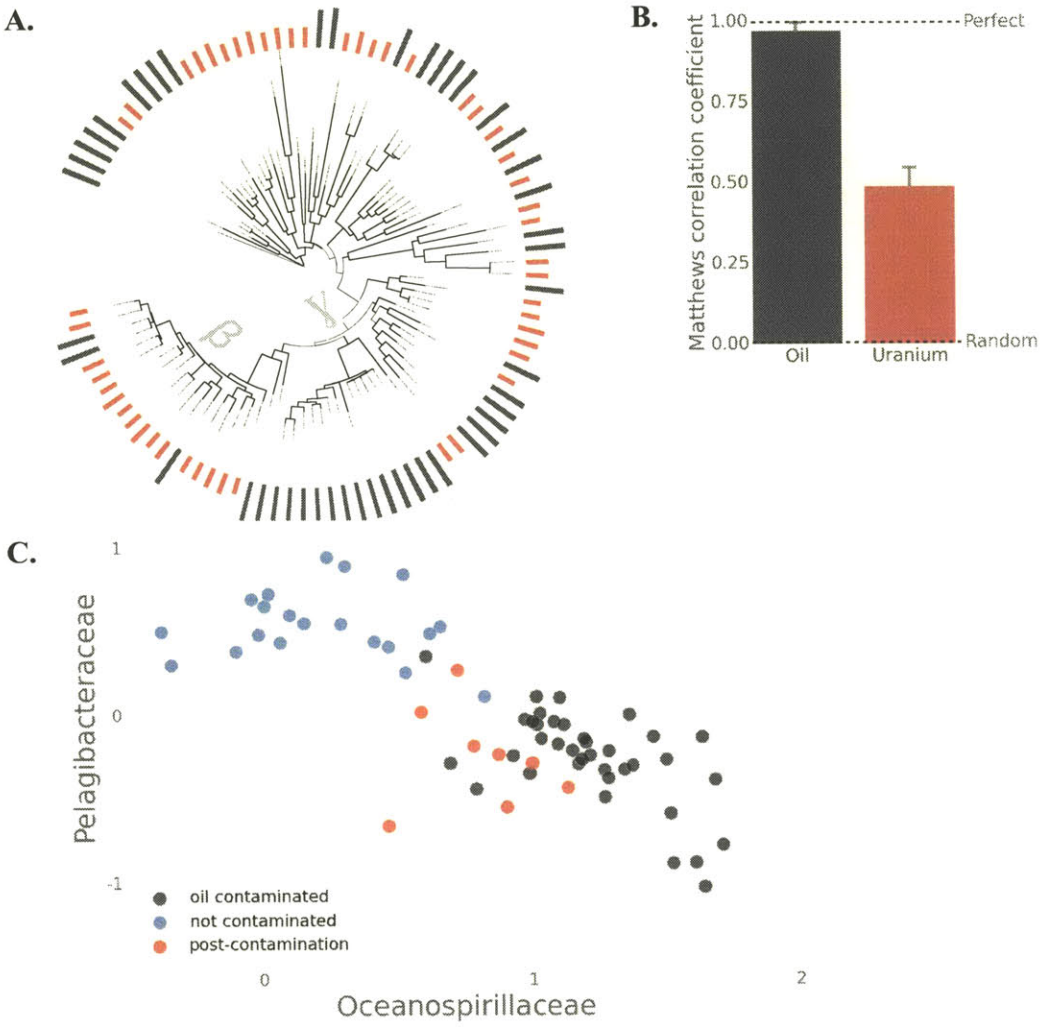
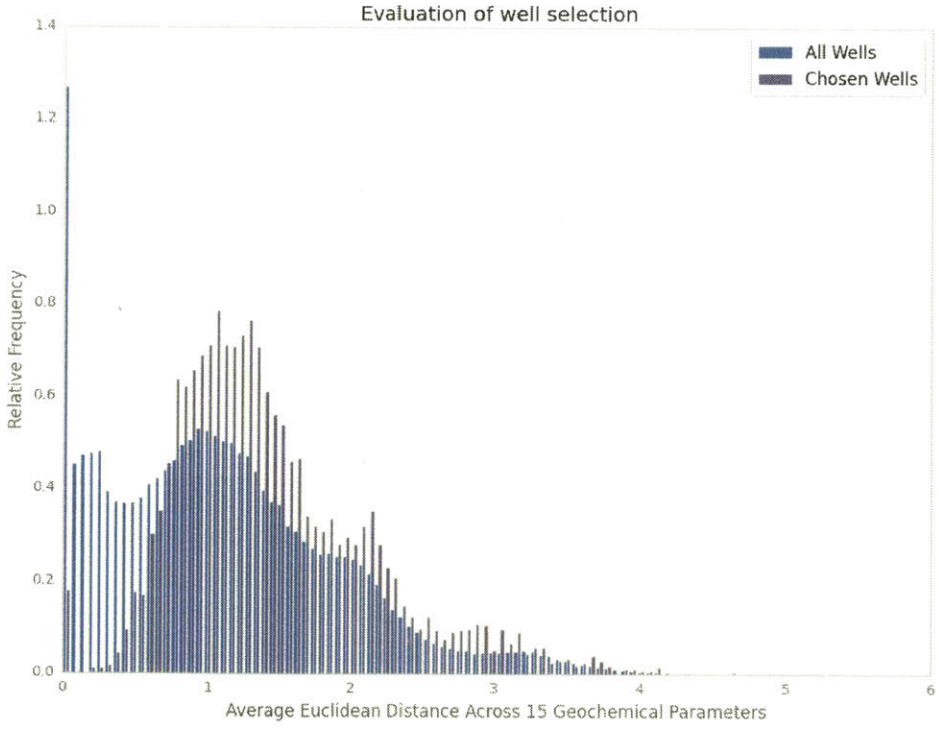


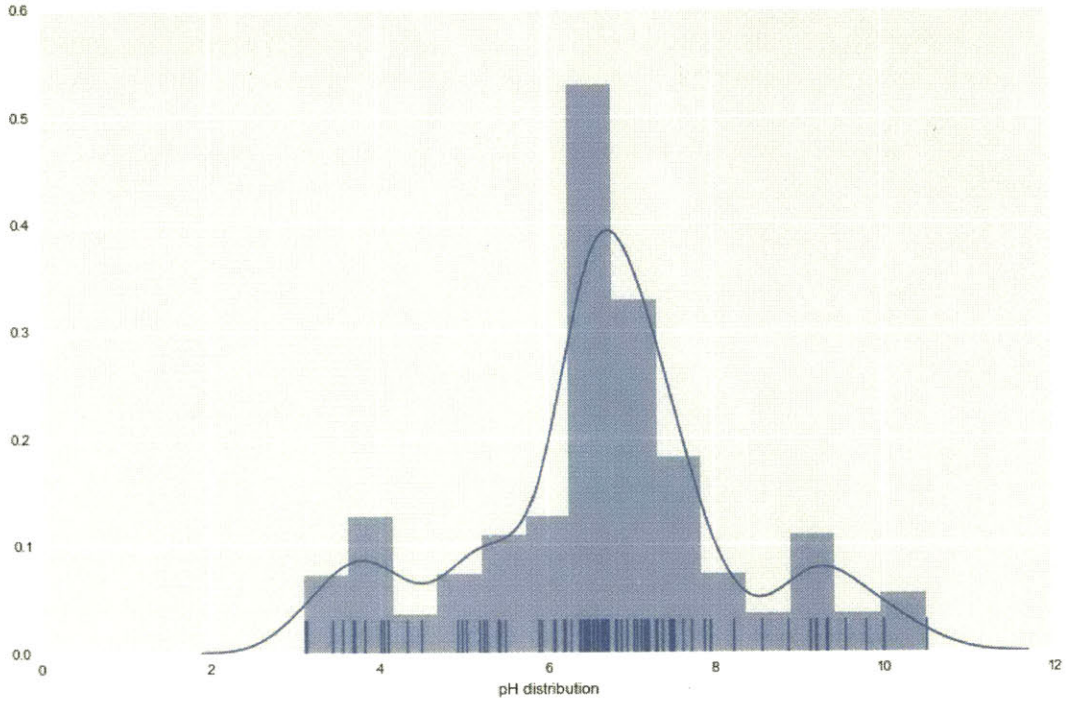
Fig. 3.4: Random forests identify highly discriminative, biologically meaningful taxonomic groups that predict environmental conditions. To understand the remarkable performance of the oil classifier, we have plotted a phylogenetic tree (**a**) that includes the 50 most informative taxonomic groups for predicting uranium (red) and oil (black). The betaproteobacteria (β) and gammaproteobacteria (γ) clades are indicated. 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 nearly perfect classifiers individually (mean MCC = 0.97). Error bars for the summary of these single-feature classifiers (**b**) reflect one standard deviation. The relative abundance of two highly informative features are shown for each sample (**c**). The relative abundance is expressed as the z-score of each group relative to the abundances of other taxonomic groups from the same sample.

3.5 Supplemental Figures and Tables

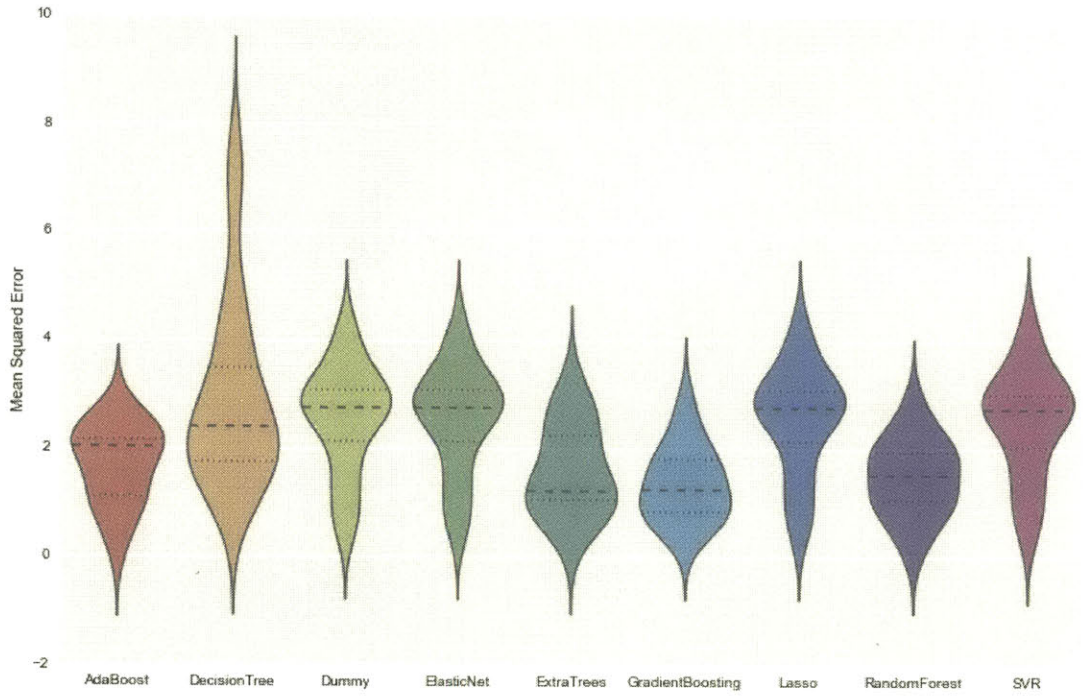
Supplemental Fig. 3.1 - Studied wells reflect a diverse subset of all available wells The pair-wise geochemical dissimilarity of wells selected for inclusion in the study (purple) and all wells with geochemical data available at Oak Ridge (blue). We computed the Euclidean distance across all of the available geochemical parameters for each pair of wells with data available (n= 834) using the most recent available historical data. Geochemical features were normalized to create a unitless metric. The resulting units for the Euclidean distance are arbitrary. The y-axis indicates the relative frequency of counts from each category. Counts have been normalized so that both categories sum to an equal number. As desired, 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, $p < 1e\text{-}10$, mann-whitney u-test).



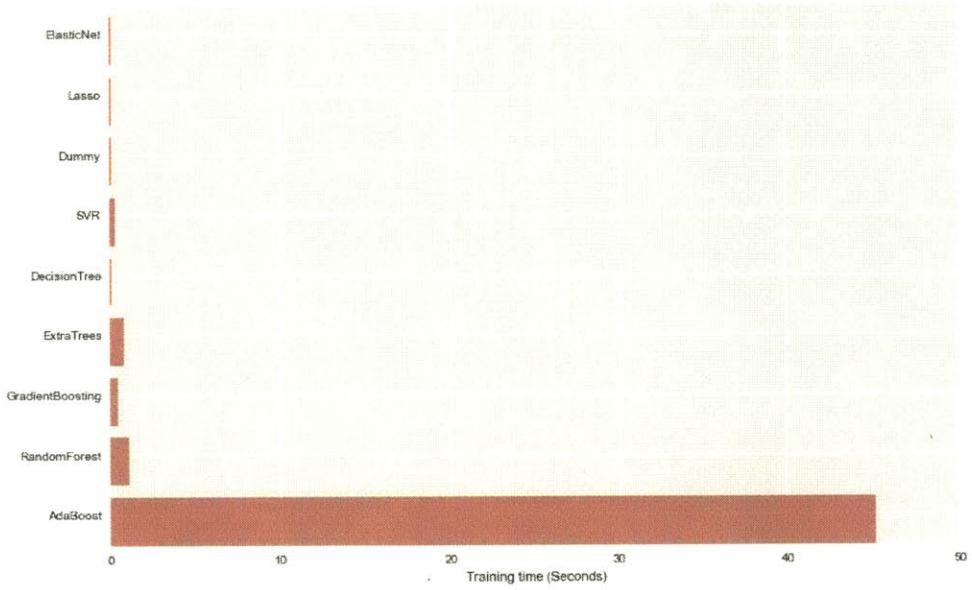
Supplemental Fig. 3.2 - Distribution of target parameter
The distribution of pH values across wells sampled at the Oak Ridge Field Site. This distribution is centered about a normal pH, however, there are several highly basic and highly acidic outliers within the dataset. Prediction of pH values from this dataset using 16S rRNA training data was used for algorithm benchmarking and selection.



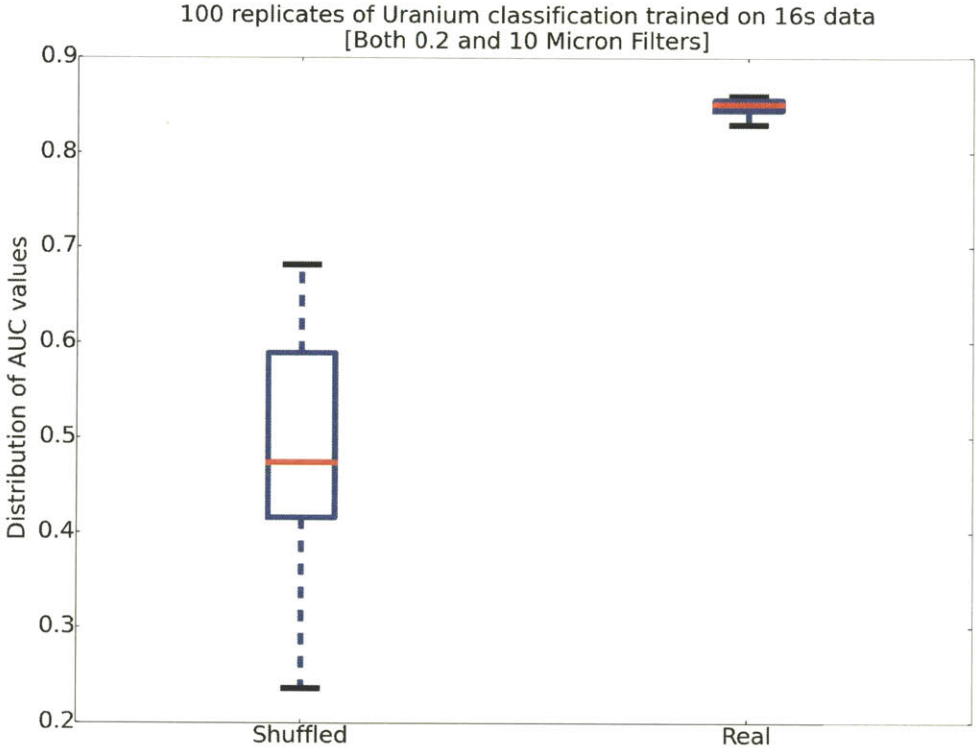
Supplemental Fig. 3.3 - Algorithm performance for regression of test dataset
We evaluated 9 popular machine learning algorithms against our pH regression benchmark test. Below we plot the distribution of mean-squared errors for the prediction of pH values for each well from each of the algorithms. We have shown the distribution of errors for each validation set in a ten-fold cross validation. In the violin plots below, the inferred distribution of errors for each model is plotted, with the median and inner-quartile range marked by the heavy and light dashed line respectively. We found that the three decision-tree based ensemble methods (RandomForest, ExtraTrees and GradientBoosting) achieved the best performance.



Supplemental Fig. 3.4 - Comparison of algorithm training speed
An additional criterion for algorithm selection was the practical constraint of training time. Here we plot the training time on a relatively small 16S rRNA dataset to predict a single geochemical parameter (pH), however for practical use it was important to select an algorithm that could be re-trained many times independently to predict many geochemical parameters or perform extensive cross-validation. We found that AdaBoost was an slow outlier, but that all other algorithms were practical to quickly run in parallel.

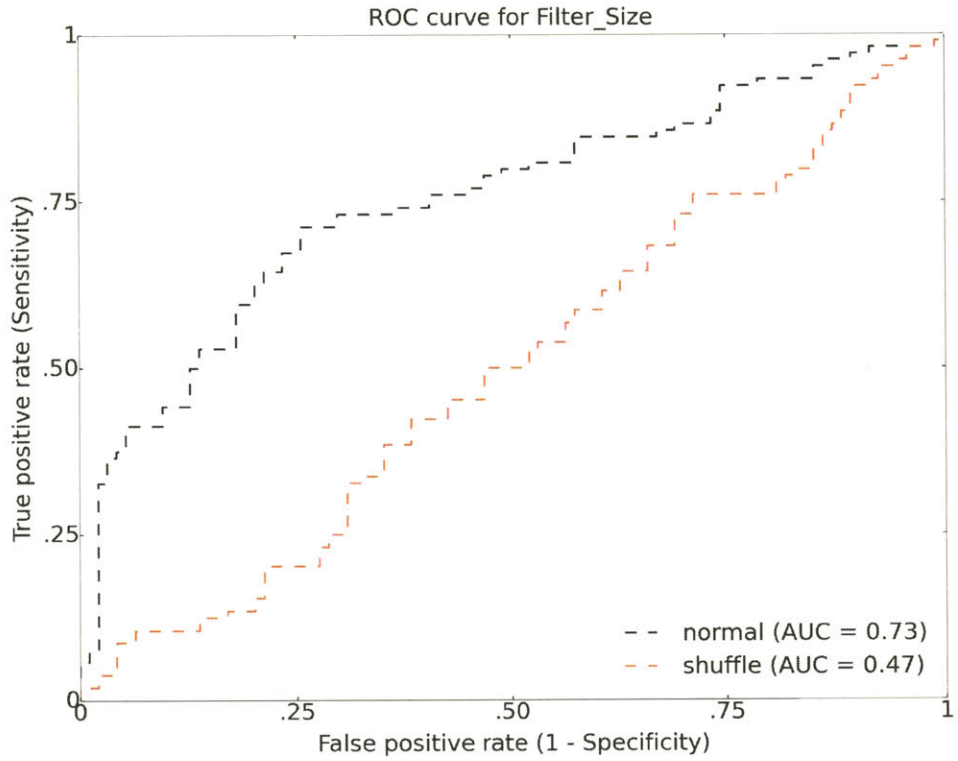


Supplemental Fig. 3.5 - Permutation test validation
The area under the receiver operator curve (AUC) was computed for each of 100 independently trained Random forests using either shuffled or real (unshuffled) data. The distribution of AUC values that resulted from this experiment indicate that shuffled data behaves as a random classifier should with classification near the 0.5 mark of non-discrimination on average. Models trained on real data consistently perform as effective classifiers with a much higher mean and much lower variance.



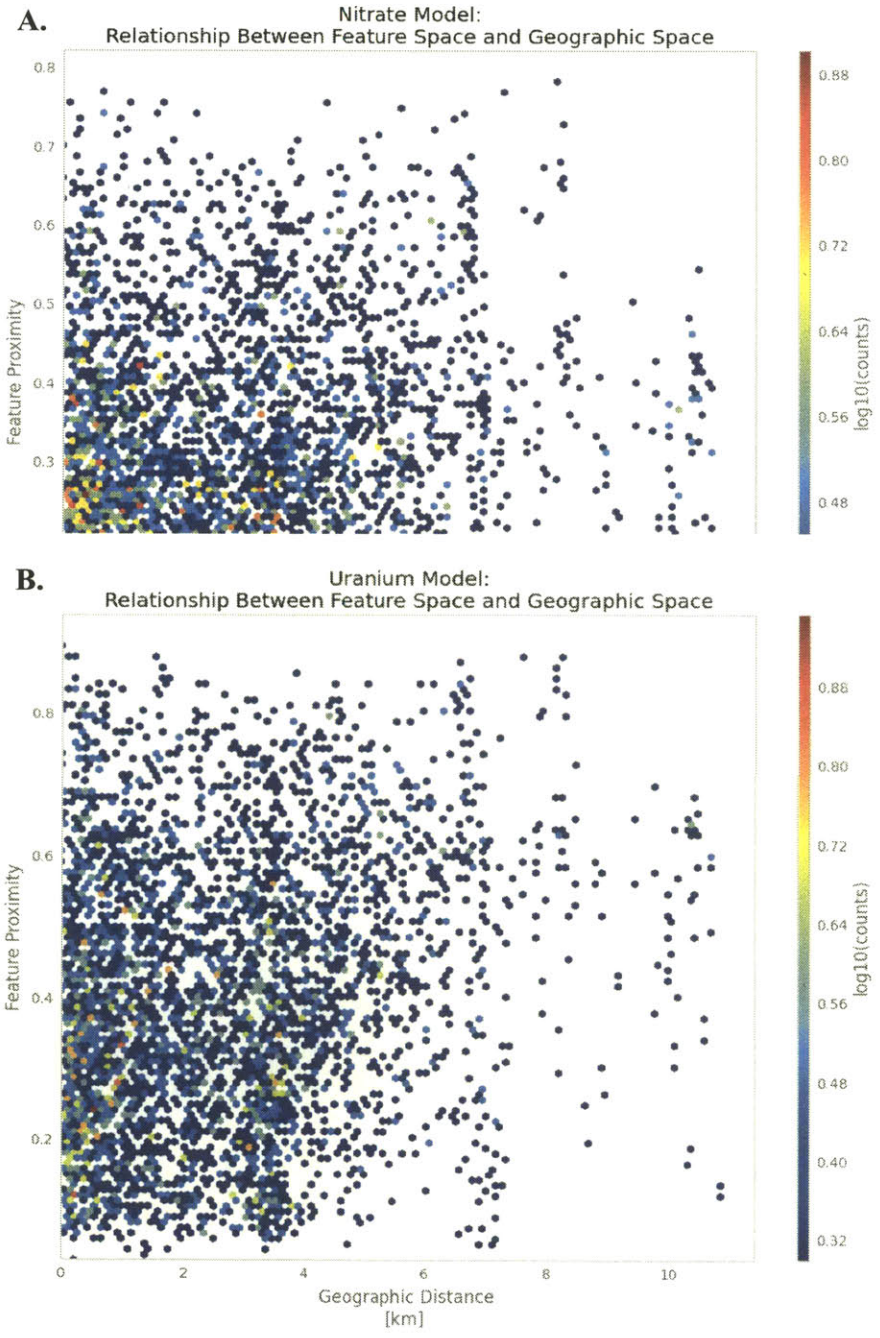
Supplemental Fig. 3.6 - Predicting filter-size with Random forest
The receive operator curve for classifying whether samples are drawn from the 10 μm or 0.2 μm filter. Because each well is sampled once at each filter size, geographic and geochemical confounds are controlled for in this analysis. The only difference between each sample is the filter size from which it is drawn. Surprisingly, we are able to predict filter size, indicating that there is discernible ecological structure between filter sizes, even when other features are controlled. The black line shows the observed ROC and the red line shows the results for a shuffled control, which approximates a random classifier.

A.

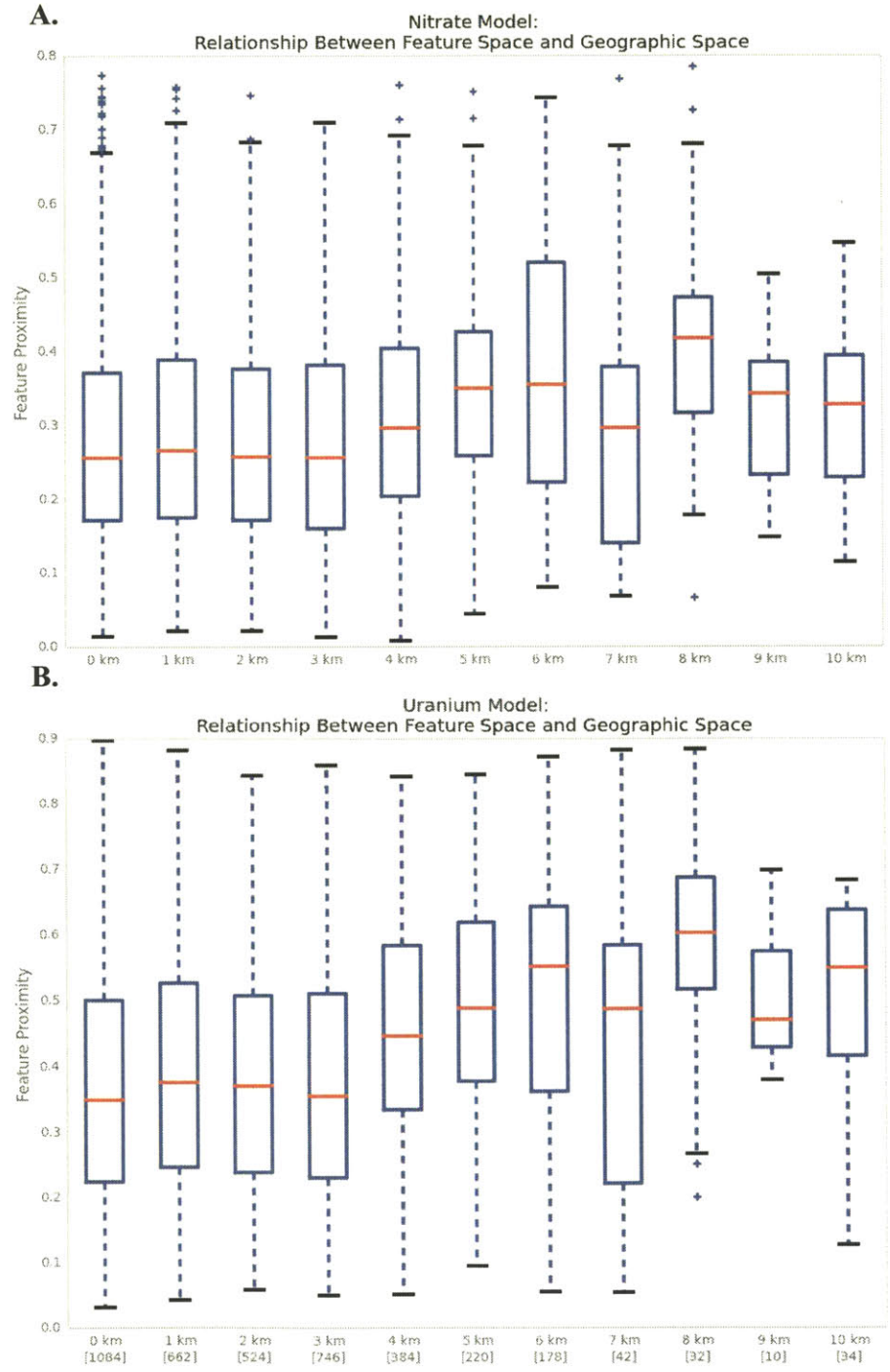


Supplemental Fig. 3.7 - Relationship between feature proximity and geographic distance

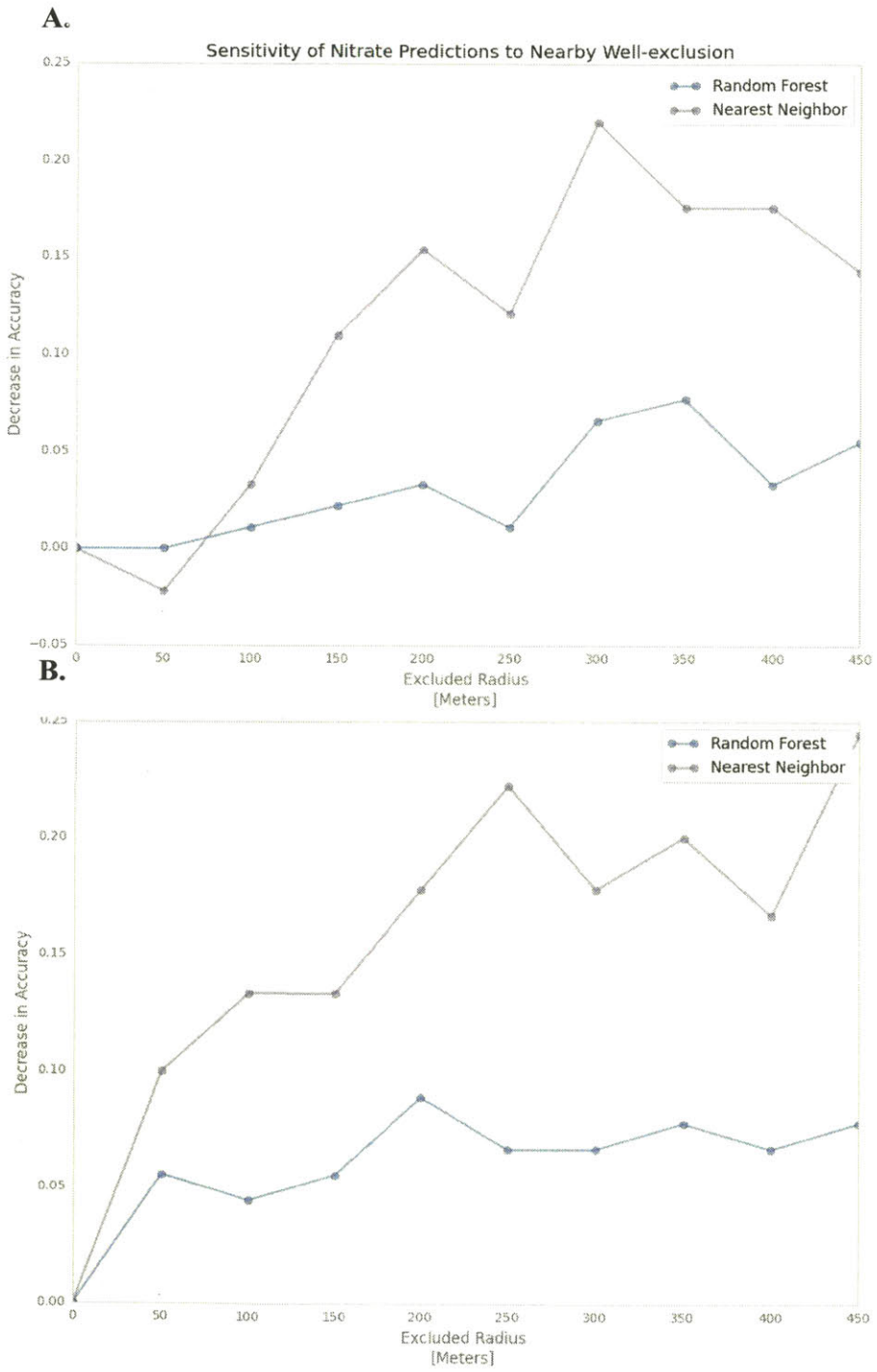
This figure illustrates the relationship between proximity in feature space and geographic space for each pair of wells at the field site for nitrate (Panel A) and Uranium (Panel B). Feature proximity (y-axis) is shown in arbitrary units, distance (x-axis) is shown in kilometers. When multiple pairs are observed with the same relationship, the data is binned to allow density to be accurately visualized. The number of data points in each bin is depicted by the color gradient on the right.



Supplemental Fig. 3.8 - Summary of feature-geography proximity relationship
As in Supplemental Fig 2.7 above, this figure plots the relationship between feature space and geographic space for nitrate (Panel A) and Uranium (Panel B). To make the data easier to visualize and interpret, here we have grouped all data into 1-km bins and plotted the distribution of feature proximity scores within each distance bin. Each boxplot shows the median (red line) inter-quartile range (box) and whiskers are 1.5 times the inter-quartile range. The number of observations in each bin is reported below each bin label on the x-axis.



Supplemental Fig. 3.9 - Geographic sensitivity of Random forest
Comparison of the impact of geographic exclusion on classification accuracy for Random forest and the nearest neighbor model. We performed leave one out cross validation of each well, varying the training set by excluding wells in an increasing radius ranging from 0 to 450 meters around each well. This well exclusion reduced prediction accuracy for both models, but had a much more dramatic impact on nearest neighbor. The decrease in accuracy relative to the base-case where no wells are excluded from the training set is shown in the y axis.



Supplemental Table 3.1: Selection of K for K-nearest neighbors model

Accuracy of classification for k-nearest neighbor models				
	K=1	K=2	K=4	K=8
Uranium	**0.86**	0.77	0.77	0.84
Nitrate	**0.77**	0.68	0.69	0.73

Conclusions and future directions

This work integrates comparative genomics and metagenomics with emerging computational tools to present a view of bacterial communities as connected networks that are highly responsive to the environment. The network of HGT presented in the first chapter illustrates long-term adaptation to the environment through genetic evolution. The analysis of CRISPR arrays presented in the second chapter provides insight into the breadth of the network of mobile genetic elements that tie together bacterial communities and enable this HGT. The ability to accurately predict geochemistry using 16S rRNA sequence data in the final chapter demonstrates a shorter-term mechanism for environmental interactions with bacterial networks, as the relative abundances of characteristic strains vary in accordance with environmental conditions. This work provides a number of important contributions to the field. Of equal importance to the questions resolved in this work are the new questions that are opened, providing many exciting directions for future investigators to explore.

The first chapter demonstrates that HGT is significantly enriched among bacteria isolated from similar environments. As the definition of ecological overlap is narrowed, for example, from human-associated bacteria to gut-associated bacteria to gut-associated pathogens, the frequency of observed HGT increases. I propose that each HGT event reflects the product of transfer and subsequent selection. Strains with greater ecological overlap are more likely to share genes that are mutually beneficial and undergo a selective sweep. However strains from the same environment are also more likely to physically interact, providing more opportunities for transfer. Examination of the historical record encoded in bacterial genomes is insufficient to distinguish between these two mechanisms that could underlie the observed ecological enrichment.

Future work could focus on developing an experimental system for characterizing HGT in real time to distinguish between these effects. For example, a fluorescently labeled strain could be introduced into a controlled environment, with the fluorescent protein linked to a gene for degrading a novel carbohydrate like laminarin. The rate of HGT could be quantitatively measured using flow-seq to perform 16S rRNA sequencing on

cells with or without the fluorescent label. Importantly, the selective advantage of the introduced gene could be systematically explored by varying the concentration of laminarin and glucose in the media followed by observation of the frequency with which cells at different phylogenetic distances acquire the gene. Under conditions of high laminarin and low glucose, cells that acquire the gene would have a significant advantage and the gene would be expected to rapidly sweep through a population after transfer. Conversely under conditions of high glucose and no laminarin, there would not be a selective advantage and the rate of HGT would reflect the 'neutral' rate in the absence of positive selection. This basic experimental design is one approach that future investigators could pursue in order to distinguish between elevated transfer or subsequent selection as drivers of ecological enrichment in HGT.

The evolutionary rate heuristic employed in chapter one does not distinguish between the mechanisms by which genes are transferred. The relative importance of phage and plasmids in facilitating HGT remains unclear. An experimental system similar to that described above could be used to measure the promiscuity of individual phage and plasmids in the specific community that is being evaluated, however the general importance of MGE is difficult to assess. It is also possible that many long distance gene transfers are transmitted between intermediates, flowing among both phage and plasmids before arriving in the genome where they are observed. Determining the relative importance of the different mechanisms of HGT remains an open question that future efforts should address.

Another intriguing area for research is clarifying the relevant temporal scales that HGT occurs under. Chapter one presents evidence of surprisingly recent transfers of nearly identical DNA. Although a molecular clock could be used to place bounds on the temporal limits of these events, it would be difficult to calibrate such a clock as these transferred genes are unlikely to conform to the substitution rates experienced by the core genes typically used for these calculations. It is possible that historical events, such as the introduction of synthetic antibiotics, could be used to calibrate the rate of HGT in some

cases, but given rapid rates of sequence evolution in MGE it would be difficult to generalize from such an observation.

An alternative approach would be to search for enriched rates of transfer among relatively isolated environments. Individual humans can be seen as ecological islands, with limited migration between them but strong interactions within individuals. If the same ecological enrichment identified at a broad scope in chapter one is recovered within a single individual, this would provide a narrow temporal constraint on ecological HGT. An ideal experiment would evaluate rates of transfer within and between twins that presumably inherited a similar community at birth. As a result, distinct HGTs should be the product of evolution during the individual's life. If there is more HGT within an individual than between individuals, this would suggest HGT can fix in a natural community on the scale of a human lifetime. Ilana Brito is currently pursuing this approach to evaluate the temporal scope of ecologically enriched HGT.

The analysis in chapter one focuses on the transfer of protein-coding DNA and was explicitly checked for open reading frames (>99% of transfers contain an ORF). However, an overlooked consequence of this ubiquitous exchange of DNA is that non-coding regions might also be subject to transfer and recombination, enabling rapid rewiring of regulatory networks. Indeed, in a separate work beyond the scope of this thesis, I have collaborated with Yaara Oren to demonstrate that in addition to HGT, horizontal regulatory transfer (HRT) is also ubiquitous, occurring across the bacterial domain (Oren et al. 2014). The ability to tap a broad pool of regulatory sequences suggests that in addition to an environment specific meta-genome, there is an unexplored parallel pool of sequences, the meta-regulome, which bacteria use to rapidly alter their gene expression in response to environmental change.

The second chapter uses CRISPR arrays to explore the host range of mobile genetic elements (MGE). The discovery of identical spacers in distantly related strains strongly implies broad host range MGE. Although I show that many of the spacers match sequenced phage and plasmid genomes, many genes are commonly shared among phage,

plasmids and their hosts, complicating these interpretations. Future work should aim to distinguish more clearly between phage and plasmid targets among identical spacer repeats.

Functional characterization of DNA targeted by these spacers will be another interesting avenue for further investigation. Given the preponderance of hypothetical and poorly annotated genes in MGE, completing this analysis will require careful annotation and review. However, the results would provide the first view into the types of genes that are most frequently targeted by CRISPR.

This chapter is intended to explore CRISPR arrays to critically evaluate long-standing dogma on the host range of mobile elements, however the resulting network should be a fruitful topic for future research. For example, it will now be possible to determine whether there are hubs in this network of shared CRISPR or what features influence the probability of sharing CRISPR spacers. Given the ecological structure of gene exchange discussed in chapter one, it will be intriguing for future efforts to determine whether the same principles apply to CRISPR arrays.

I find that matching CRISPR spacers are also themselves, the product of frequent HGT. This assertion is supported by the identification of identical CRISPR repeats among genomes with matching spacers. It will be important to perform rigorous phylogenetic confirmation of this observation in the future. This can be done through phylogenetic reconstruction of the Cas genes to provide an additional line of evidence supporting transfer. This analysis will also help cast light on the ambiguous cases that share similar, but not identical repeats. Verifying the evolutionary origins of putatively transferred CRISPR arrays will be a valuable topic for follow-up work.

The final chapter introduces the concept of using native bacterial communities as biosensors to measure environmental features. This work is intended as a proof-of-principle, rather than a practical implementation. With evidence supporting the utility of this basic approach, future efforts can focus on refining and improving this method.

To maximize the probability of detecting meaningful geochemical signals in metagenomic data, I have focused these efforts on the extreme gradients created at a nuclear waste site and at one of the largest oil spills in US history. Although these extreme gradients can be easily detected, future efforts will need to explore the limits of this approach. If a pH gradient of 1 unit can be detected, does the same hold true for a gradient of 0.1? It will be useful to define the practical detection limits of indigenous biosensors.

One of the most intriguing observations presented in chapter three is the discovery that previously contaminated ocean sites can be identified through metagenomic analysis even after oil contamination has disappeared. However, it is unclear how long this ecological memory might persist and how robust it might be given the modest sample size explored in this chapter. It would be of great interest to do a longitudinal study of a transient chemical manipulation to determine the duration of this memory effect. Scott Oleson is now pursuing this question by exploring microbial signatures and metabolic responses in amended wells at the Oak Ridge Field Site used in this study.

Although the successful application of the indigenous biosensor approach to two distinct environments suggests that the method may be generalizable, it will require much more extensive evaluation to determine what environments might be amenable to this technique. I expect that the key considerations for success will be the degree of environmental distinction (e.g. are the environments sufficiently different to create a detectible signal) and the number of independent training samples (a larger training set will enable the detection of smaller signals).

Many new features beyond 16S rRNA should be explored as features for classifying environmental samples. For example, shotgun metagenomics would provide functional insight into the community, while metatranscriptomics could provide greater temporal resolution. Even for applications with 16S rRNA data, the relationship between sequencing depth and performance needs to be optimized.

Of additional practical importance, the relationship between the size of the training set and model performance will be critical for practical deployment of indigenous biosensors. Generally, model performance is expected to improve as the training set expands, but it is unclear what the slope and shape of this curve will be. Given that the cost and complexity of implementing this approach will depend largely on the number of samples included, this will be an important relationship to understand.

In both the ocean and groundwater examples explored in chapter three, samples were drawn from a relatively narrow geographic range. It is unclear whether it will be possible to, for example, extrapolate a classifier trained on data from Oak Ridge to other nuclear waste sites. Given the remarkable diversity observed at Oak Ridge, it seems likely that many OTUs will be unique to each site, precluding the extension of a model from one site to another. Evaluating the ability to generalize beyond the location of the training set will be an important task for future efforts.

As new technologies continue to emerge, facilitating ever cheaper and easier generation of genomic data, thoughtful computational analysis will continue to become an increasingly critical tool for effective research. This work has focused on the development and deployment of computational approaches to explore the relationships between bacterial communities and their environments. Although this work explores many aspects of bacterial ecology and evolution, it is unified by the view that bacterial systems are best understood as interacting networks that are shaped by the environments in which they reside.

Bibliography

Altschul, S. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1006/jmbi.1990.9999.

Aravind, L, Roman L Tatusov, Yuri I Wolf, D.Roland Walker, and Eugene V Koonin. 1998. "Evidence for Massive Gene Exchange between Archaeal and Bacterial Hyperthermophiles." *Trends in Genetics* 14 (11): 442–44. doi:10.1016/S0168-9525(98)01553-4.

Arigoni, Fabrizio, Francois Talabot, Manuel Peitsch, Michael D. Edgerton, Eric Meldrum, Elisabeth Allet, Richard Fish, Therese Jamotte, Marie-Laure Curchod, and Hannes Loferer. 1998. "A Genome-Based Approach for the Identification of Essential Bacterial Genes." *Nature Biotechnology* 16 (9): 851–56. doi:10.1038/nbt0998-851.

Belkin, Shimshon. 2003. "Microbial Whole-Cell Sensing Systems of Environmental Pollutants." *Current Opinion in Microbiology* 6 (3): 206–12. doi:10.1016/S1369-5274(03)00059-6.

Benson, Dennis A., Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler. 2005. "GenBank." *Nucleic Acids Research* 33 (Database issue): D34–D38. doi:10.1093/nar/gki063.

Bland, Charles, Teresa L. Ramsey, Fareedah Sabree, Micheal Lowe, Kyndall Brown, Nikos C. Kyrpides, and Philip Hugenholtz. 2007. "CRISPR Recognition Tool (CRT): A Tool for Automatic Detection of Clustered Regularly Interspaced Palindromic Repeats." *BMC Bioinformatics* 8 (1): 209. doi:10.1186/1471-2105-8-209.

Blattner, Frederick R., Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, et al. 1997. "The Complete Genome Sequence of Escherichia Coli K-12." *Science* 277 (5331): 1453–62. doi:10.1126/science.277.5331.1453.

Boucher, Y., O. X. Cordero, A. Takemura, D. E. Hunt, K. Schliep, E. Bapteste, P. Lopez, C. L. Tarr, and M. F. Polz. 2011. "Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global Vibrio Cholerae Populations." *mBio* 2 (2): e00335–10–e00335–10. doi:10.1128/mBio.00335-10.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5): 335–36. doi:10.1038/nmeth.f.303.

Caporaso, J. Gregory, Kyle Bittinger, Frederic D. Bushman, Todd Z. DeSantis, Gary L. Andersen, and Rob Knight. 2010. "PyNAST: A Flexible Tool for Aligning

Sequences to a Template Alignment." *Bioinformatics (Oxford, England)* 26 (2): 266–67. doi:10.1093/bioinformatics/btp636.

Caporaso, J. Gregory, Christian L. Lauber, William A. Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M. Owens, et al. 2012. "Ultra-High-Throughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms." *The ISME Journal* 6 (8): 1621–24. doi:10.1038/ismej.2012.8.

Caporaso, JG, DN Miller, EL Bryant, and DNA. 1999. "Et Al." *Ultrahighthroughput Microbial Community Analysis on the Illumina HiSeq and MiSeq Platforms Isme Journal 6 J E Madsen and W C Ghiorse Evaluation and Optimization of and Purification Procedures for Soil and Sediment Samples Applied and Environmental Microbiology 47154724* 80 (12 SRC - GoogleScholar): 1621–24.

Carini, Paul, Laura Steindler, Sara Beszteri, and Stephen J Giovannoni. 2013. "Nutrient Requirements for Growth of the Extreme Oligotroph 'Candidatus Pelagibacter Ubique' HTCC1062 on a Defined Medium." *The ISME Journal* 7 (3): 592–602. doi:10.1038/ismej.2012.122.

Caro-Quintero, Alejandro, Jie Deng, Jennifer Auchtung, Ingrid Brettar, Manfred G Höfle, Joel Klappenbach, and Konstantinos T Konstantinidis. 2011. "Unprecedented Levels of Horizontal Gene Transfer among Spatially Co-Occurring Shewanella Bacteria from the Baltic Sea." *The ISME Journal* 5 (1): 131–40. doi:10.1038/ismej.2010.93.

Chen, J., and R. P. Novick. 2009. "Phage-Mediated Intergeneric Transfer of Toxin Genes." *Science* 323 (5910): 139–41. doi:10.1126/science.1164783.

Clatworthy, Anne E, Emily Pierson, and Deborah T Hung. 2007. "Targeting Virulence: A New Paradigm for Antimicrobial Therapy." *Nature Chemical Biology* 3 (9): 541–48. doi:10.1038/nchembio.2007.24.

Cole, J R, B Chai, R J Farris, Q Wang, S A Kulam, D M McGarrell, G M Garrity, and J M Tiedje. 2005. "The Ribosomal Database Project (RDP-II): Sequences and Tools for High-Throughput rRNA Analysis." *Nucleic Acids Research* 33 (Database issue): D294–296. doi:10.1093/nar/gki038.

D'Souza, S. F. 2001. "Microbial Biosensors." *Biosensors and Bioelectronics* 16 (6): 337–53. doi:10.1016/S0956-5663(01)00125-7.

Darwin, Charles. 1859. *On the Origin of the Species by Natural Selection.* Murray.

Daubin, Vincent, and Howard Ochman. 2004. "Quartet Mapping and the Extent of Lateral Transfer in Bacterial Genomes." *Molecular Biology and Evolution* 21 (1): 86–89. doi:10.1093/molbev/msg234.

David, Lawrence A., and Eric J. Alm. 2011. "Rapid Evolutionary Innovation during an Archaean Genetic Expansion." *Nature* 469 (7328): 93–96. doi:10.1038/nature09649.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006a. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. doi:10.1128/AEM.03006-05.

———. 2006b. "Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. doi:10.1128/AEM.03006-05.

Dubey, Gyanendra P., and Sigal Ben-Yehuda. 2011. "Intercellular Nanotubes Mediate Bacterial Communication." *Cell* 144 (4): 590–600. doi:10.1016/j.cell.2011.01.015.

Edgar, R. C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61. doi:10.1093/bioinformatics/btq461.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster than BLAST." *Bioinformatics* 26 (19): 2460–61. doi:10.1093/bioinformatics/btq461.

Fischer, Thorsten, Ashutosh Agarwal, and Henry Hess. 2009. "A Smart Dust Biosensor Powered by Kinesin Motors." *Nature Nanotechnology* 4 (3): 162–66. doi:10.1038/nnano.2008.393.

Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, et al. 2011. "Ensembl 2011." *Nucleic Acids Research* 39 (Database issue): D800–806. doi:10.1093/nar/gkq1064.

Frost, Laura S., Raphael Leplae, Anne O. Summers, and Ariane Toussaint. 2005. "Mobile Genetic Elements: The Agents of Open Source Evolution." *Nature Reviews Microbiology* 3 (9): 722–32. doi:10.1038/nrmicro1235.

Garg, Pallavi, Antonia Aydanian, David Smith, Morris J Glenn, G. Balakrish Nair, and O. Colin Stine. 2003. "Molecular Epidemiology of O139 Vibrio Cholerae: Mutation, Lateral Gene Transfer, and Founder Flush." *Emerging Infectious Diseases* 9 (7): 810–14. doi:10.3201/eid0907.030038.

Gianoulis, Tara A., Jeroen Raes, Prianka V. Patel, Robert Bjornson, Jan O. Korbel, Ivica Letunic, Takuji Yamada, et al. 2009. "Quantifying Environmental Adaptation of Metabolic Pathways in Metagenomics." *Proceedings of the National Academy of Sciences* 106 (5): 1374–79. doi:10.1073/pnas.0808022106.

Gill, S. R. 2006. "Metagenomic Analysis of the Human Distal Gut Microbiome." *Science* 312 (5778): 1355–59. doi:10.1126/science.1124234.

Gogarten, J. Peter, W. Ford Doolittle, and Jeffrey G. Lawrence. 2002. "Prokaryotic Evolution in Light of Gene Transfer." *Molecular Biology and Evolution* 19 (12): 2226–38.

Gogarten, J. Peter, and Jeffrey P. Townsend. 2005. "Horizontal Gene Transfer, Genome Innovation and Evolution." *Nature Reviews Microbiology* 3 (9): 679–87. doi:10.1038/nrmicro1204.

Green, Stefan J, Om Prakash, Puja Jasrotia, Will A Overholt, Erick Cardenas, Daniela Hubbard, James M Tiedje, et al. 2012. "Denitrifying Bacteria from the Genus Rhodanobacter Dominate Bacterial Communities in the Highly Contaminated Subsurface of a Nuclear Legacy Waste Site." *Applied and Environmental Microbiology* 78 (4): 1039–47. doi:10.1128/AEM.06435-11.

Green, Stefan J., Om Prakash, Thomas M. Gihring, Denise M. Akob, Puja Jasrotia, Philip M. Jardine, David B. Watson, Steven D. Brown, Anthony V. Palumbo, and Joel E. Kostka. 2010. "Denitrifying Bacteria Isolated from Terrestrial Subsurface Sediments Exposed to Mixed-Waste Contamination." *Applied and Environmental Microbiology* 76 (10): 3244–54. doi:10.1128/AEM.03069-09.

Hambly, Emma, and Curtis A Suttle. 2005. "The Viriosphere, Diversity, and Genetic Exchange within Phage Communities." *Current Opinion in Microbiology*, Host--microbe interactions: fungi / edited by Howard Bussey · Host--microbe interactions: parasites / edited by Artur Scherf · Host--microbe interactions: viruses / edited by Margaret CM Smith, 8 (4): 444–50. doi:10.1016/j.mib.2005.06.005.

Hazen, Terry C., Eric A. Dubinsky, Todd Z. DeSantis, Gary L. Andersen, Yvette M. Piceno, Navjeet Singh, Janet K. Jansson, et al. 2010. "Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria." *Science* 330 (6001): 204–8. doi:10.1126/science.1195979.

Hehemann, Jan-Hendrik, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, and Gurvan Michel. 2010. "Transfer of Carbohydrate-Active Enzymes from Marine Bacteria to Japanese Gut Microbiota." *Nature* 464 (7290): 908–12. doi:10.1038/nature08937.

Hendrix, Roger W., Margaret C. M. Smith, R. Neil Burns, Michael E. Ford, and Graham F. Hatfull. 1999. "Evolutionary Relationships among Diverse Bacteriophages and Prophages: All the World's a Phage." *Proceedings of the National Academy of Sciences* 96 (5): 2192–97. doi:10.1073/pnas.96.5.2192.

Himmelreich, Ralf, Helmut Hilbert, Helga Plagens, Elsbeth Pirkl, Bi-Chen Li, and Richard Herrmann. 1996. "Complete Sequence Analysis of the Genome of the Bacterium Mycoplasma Pneumoniae." *Nucleic Acids Research* 24 (22): 4420–49. doi:10.1093/nar/24.22.4420.

Hyman, Paul, and Stephen T. Abedon. 2010. "Bacteriophage Host Range and Bacterial Resistance." *Advances in Applied Microbiology* 70: 217–48. doi:10.1016/S0065-2164(10)70007-1.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. "Horizontal Gene Transfer among Genomes: The Complexity Hypothesis." *Proceedings of the National Academy of Sciences* 96 (7): 3801–6. doi:10.1073/pnas.96.7.3801.

Kavitha, S., R. Selvakumar, M. Sathishkumar, K. Swaminathan, P. Lakshmanaperumalsamy, A. Singh, and S. K. Jain. 2009. "Nitrate Removal Using Brevundimonas Diminuta MTCC 8486 from Ground Water." *Water Science & Technology* 60 (2): 517. doi:10.2166/wst.2009.378.

Kim, Kwang Sik. 2003. "Neurological Diseases: Pathogenesis of Bacterial Meningitis: From Bacteraemia to Neuronal Injury." *Nature Reviews Neuroscience* 4 (5): 376–85. doi:10.1038/nrn1103.

Koonin, E. V., L. Aravind, and A. S. Kondrashov. 2000. "The Impact of Comparative Genomics on Our Understanding of Evolution." *Cell* 101 (6): 573–76.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. "Horizontal Gene Transfer in Prokaryotes: Quantification and Classification." *Annual Review of Microbiology* 55: 709–42. doi:10.1146/annurev.micro.55.1.709.

Kumarasamy, Karthikeyan K, Mark A Toleman, Timothy R Walsh, Jay Bagaria, Fafhana Butt, Ravikumar Balakrishnan, Uma Chaudhary, et al. 2010. "Emergence of a New Antibiotic Resistance Mechanism in India, Pakistan, and the UK: A Molecular, Biological, and Epidemiological Study." *The Lancet Infectious Diseases* 10 (9): 597–602. doi:10.1016/S1473-3099(10)70143-2.

Kunin, Victor, Rotem Sorek, and Philip Hugenholtz. 2007. "Evolutionary Conservation of Sequence and Secondary Structures in CRISPR Repeats." *Genome Biology* 8 (4): R61. doi:10.1186/gb-2007-8-4-r61.

Lauber, Christian L., Micah Hamady, Rob Knight, and Noah Fierer. 2009. "Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale." *Applied and Environmental Microbiology* 75 (15): 5111–20. doi:10.1128/AEM.00335-09.

Lawrence, Jeffrey G., and Heather Hendrickson. 2003. "Lateral Gene Transfer: When Will Adolescence End?: Growing Pains for Gene Transfer." *Molecular Microbiology* 50 (3): 739–49. doi:10.1046/j.1365-2958.2003.03778.x.

Lester, C. H., N. Frimodt-Moller, T. L. Sorensen, D. L. Monnet, and A. M. Hammerum. 2006. "In Vivo Transfer of the vanA Resistance Gene from an Enterococcus Faecium Isolate of Animal Origin to an E. Faecium Isolate of Human Origin in the Intestines of Human Volunteers." *Antimicrobial Agents and Chemotherapy* 50 (2): 596–99. doi:10.1128/AAC.50.2.596-599.2006.

Ley, Ruth E., Peter J. Turnbaugh, Samuel Klein, and Jeffrey I. Gordon. 2006. "Microbial Ecology: Human Gut Microbes Associated with Obesity." *Nature* 444 (7122): 1022–23. doi:10.1038/4441022a.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Liu, B., and M. Pop. 2009. "ARDB--Antibiotic Resistance Genes Database." *Nucleic Acids Research* 37 (Database): D443–D447. doi:10.1093/nar/gkn656.

Makarova, Kira S., L. Aravind, Michael Y. Galperin, Nick V. Grishin, Roman L. Tatusov, Yuri I. Wolf, and Eugene V. Koonin. 1999. "Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell." *Genome Research* 9 (7): 608–28. doi:10.1101/gr.9.7.608.

Markowitz, V. M. 2006. "The Integrated Microbial Genomes (IMG) System." *Nucleic Acids Research* 34 (90001): D344–D348. doi:10.1093/nar/gkj024.

Mazodier, P, and J Davies. 1991. "Gene Transfer Between Distantly Related Bacteria." *Annual Review of Genetics* 25 (1): 147–71. doi:10.1146/annurev.ge.25.120191.001051.

McClelland, Michael, Kenneth E. Sanderson, John Spieth, Sandra W. Clifton, Phil Latreille, Laura Courtney, Steffen Porwollik, et al. 2001. "Complete Genome Sequence of Salmonella Enterica Serovar Typhimurium LT2." *Nature* 413 (6858): 852–56. doi:10.1038/35101614.

Metcalf, J. L., L. Wegener Parfrey, A. Gonzalez, C. L. Lauber, D. Knights, G. Ackermann, G. C. Humphrey, et al. 2013. "A Microbial Clock Provides an Accurate Estimate of the Postmortem Interval in a Mouse Model System." *eLife* 2 (0): e01104–e01104. doi:10.7554/eLife.01104.

Miller, Melissa B., and Bonnie L. Bassler. 2001. "Quorum Sensing in Bacteria." *Annual Review of Microbiology* 55 (1): 165–99. doi:10.1146/annurev.micro.55.1.165.

Morris, Robert M, Michael S Rappé, Stephanie A Connon, Kevin L Vergin, William A Siebold, Craig A Carlson, and Stephen J Giovannoni. 2002. "SAR11 Clade Dominates Ocean Surface Bacterioplankton Communities." *Nature* 420 (6917): 806–10. doi:10.1038/nature01240.

Moss, Richard H., Jae A. Edmonds, Kathy A. Hibbard, Martin R. Manning, Steven K. Rose, Detlef P. van Vuuren, Timothy R. Carter, et al. 2010. "The next Generation of Scenarios for Climate Change Research and Assessment." *Nature* 463 (7282): 747–56. doi:10.1038/nature08823.

Needleman, Saul B., and Christian D. Wunsch. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins."

Journal of Molecular Biology 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4.

Ochman, Howard, S. Elwyn, and N. A. Moran. 1999. "Calibrating Bacterial Evolution." *Proceedings of the National Academy of Sciences* 96 (22): 12638–43. doi:10.1073/pnas.96.22.12638.

Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. "Lateral Gene Transfer and the Nature of Bacterial Innovation." *Nature* 405 (6784): 299–304. doi:10.1038/35012500.

Ochman, Howard, and Allan C. Wilson. 1987. "Evolution in Bacteria: Evidence for a Universal Substitution Rate in Cellular Genomes." *Journal of Molecular Evolution* 26 (1-2): 74–86. doi:10.1007/BF02111283.

Oren, Yaara, Mark B. Smith, Nathan I. Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z. Ron, Jukka Corander, Harris H. Wang, Eric J. Alm, and Tal Pupko. 2014. "Transfer of Non-Coding DNA Drives Regulatory Rewiring in Bacteria." *Submitted*, July.

Ott, Stephan J., Meike Musfeldt, Uwe Ullmann, Jochen Hampe, and Stefan Schreiber. 2004. "Quantification of Intestinal Bacterial Populations by Real-Time PCR with a Universal Primer Set and Minor Groove Binder Probes: A Global Approach to the Enteric Flora." *Journal of Clinical Microbiology* 42 (6): 2566–72. doi:10.1128/JCM.42.6.2566-2572.2004.

Papa, Eliseo, Michael Docktor, Christopher Smillie, Sarah Weber, Sarah P. Preheim, Dirk Gevers, Georgia Giannoukos, et al. 2012. "Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease." *PLoS ONE* 7 (6): e39242. doi:10.1371/journal.pone.0039242.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *J. Mach. Learn. Res.* 12 (November): 2825–30.

Preheim, Sarah P., Allison R. Perrotta, Antonio M. Martin-Platero, Anika Gupta, and Eric J. Alm. 2013. "Distribution-Based Clustering: Using Ecology To Refine the Operational Taxonomic Unit." *Applied and Environmental Microbiology* 79 (21): 6593–6603. doi:10.1128/AEM.00342-13.

Rawlings, Douglas E., and Erhard Tietze. 2001. "Comparative Biology of IncQ and IncQ-Like Plasmids." *Microbiology and Molecular Biology Reviews* 65 (4): 481–96. doi:10.1128/MMBR.65.4.481-496.2001.

Rodrigue, Sébastien, Arne C. Materna, Sonia C. Timberlake, Matthew C. Blackburn, Rex R. Malmstrom, Eric J. Alm, and Sallie W. Chisholm. 2010. "Unlocking Short Read Sequencing for Metagenomics." *PLoS ONE* 5 (7): e11840. doi:10.1371/journal.pone.0011840.

Round, June L., and Sarkis K. Mazmanian. 2009. "The Gut Microbiota Shapes Intestinal Immune Responses during Health and Disease." *Nature Reviews Immunology* 9 (5): 313–23. doi:10.1038/nri2515.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23): 7537–41. doi:10.1128/AEM.01541-09.

Smillie, Chris S, Mark B Smith, Jonathan Friedman, Otto X Cordero, Lawrence A David, and Eric J Alm. 2011. "Ecology Drives a Global Network of Gene Exchange Connecting the Human Microbiome." *Nature* 480 (7376): 241–44. doi:10.1038/nature10571.

Sorek, Rotem, Victor Kunin, and Philip Hugenholtz. 2008. "CRISPR — a Widespread System That Provides Acquired Resistance against Phages in Bacteria and Archaea." *Nature Reviews Microbiology* 6 (3): 181–86. doi:10.1038/nrmicro1793.

Stanhope, M. J., A. Lupas, M. J. Italia, K. K. Koretke, C. Volker, and J. R. Brown. 2001. "Phylogenetic Analyses Do Not Support Horizontal Gene Transfers from Bacteria to Vertebrates." *Nature* 411 (6840): 940–44. doi:10.1038/35082058.

Su, Liang, Wenzhao Jia, Changjun Hou, and Yu Lei. 2011. "Microbial Biosensors: A Review." *Biosensors and Bioelectronics* 26 (5): 1788–99. doi:10.1016/j.bios.2010.09.005.

Teramoto, Maki, Motoyuki Ohuchi, Ariani Hatmanti, Yeti Darmayati, Yantyati Widyastuti, Shigeaki Harayama, and Yukiyo Fukunaga. 2011. "Oleibacter Marinus Gen. Nov., Sp. Nov., a Bacterium That Degrades Petroleum Aliphatic Hydrocarbons in a Tropical Marine Environment." *International Journal of Systematic and Evolutionary Microbiology* 61 (Pt 2): 375–80. doi:10.1099/ijs.0.018671-0.

Thomas, Christopher M., and Kaare M. Nielsen. 2005. "Mechanisms Of, and Barriers To, Horizontal Gene Transfer between Bacteria." *Nature Reviews Microbiology* 3 (9): 711–21. doi:10.1038/nrmicro1234.

Tuller, T., Y. Girshovich, Y. Sella, A. Kreimer, S. Freilich, M. Kupiec, U. Gophna, and E. Ruppin. 2011. "Association between Translation Efficiency and Horizontal Gene Transfer within Microbial Communities." *Nucleic Acids Research* 39 (11): 4743–55. doi:10.1093/nar/gkr054.

Tyson, Gene W., and Jillian F. Banfield. 2008. "Rapidly Evolving CRISPRs Implicated in Acquired Resistance of Microorganisms to Viruses." *Environmental Microbiology* 10 (1): 200–207. doi:10.1111/j.1462-2920.2007.01444.x.

Van der Oost, John, Matthijs M. Jore, Edze R. Westra, Magnus Lundgren, and Stan J. J. Brouns. 2009. "CRISPR-Based Adaptive and Heritable Immunity in Prokaryotes." *Trends in Biochemical Sciences* 34 (8): 401–7. doi:10.1016/j.tibs.2009.05.002.

Watson, DB, and MW Kostka. 2004. "J. E. Fields and P." *M Jardine The Oak Ridge Field Research Center Conceptual Model Oak Ridge National Laboratory.*

Weitz, Joshua S., Timothée Poisot, Justin R. Meyer, Cesar O. Flores, Sergi Valverde, Matthew B. Sullivan, and Michael E. Hochberg. 2013. "Phage–bacteria Infection Networks." *Trends in Microbiology* 21 (2): 82–91. doi:10.1016/j.tim.2012.11.003.

Wu, Jun, Jong Pil Park, Kevin Dooley, Donald M. Cropek, Alan C. West, and Scott Banta. 2011. "Rapid Development of New Protein Biosensors Utilizing Peptides Obtained via Phage Display." *PLoS ONE* 6 (10): e24948. doi:10.1371/journal.pone.0024948.

Xavier, R. J., and D. K. Podolsky. 2007. "Unravelling the Pathogenesis of Inflammatory Bowel Disease." *Nature* 448 (7152): 427–34. doi:10.1038/nature06005.

Xu, Jian, Michael A. Mahowald, Ruth E. Ley, Catherine A. Lozupone, Micah Hamady, Eric C. Martens, Bernard Henrissat, et al. 2007. "Evolution of Symbiotic Bacteria in the Distal Human Intestine." *PLoS Biology* 5 (7): e156. doi:10.1371/journal.pbio.0050156.