Linking Sequence to Function in Microbial Genomics

by

Sarah Jean Spencer

B.A. Biology
Washington University in St. Louis, 2009

Submitted to the Program of Computational and Systems Biology in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

at the

Massachusetts Institute of Technology

September 2017

Signature of Author...
Computational and Systems Biology Graduate Program
July 28, 2017

Certified by...
Eric Alm
Professor of Biological Engineering
Professor of Civil and Environmental Engineering
Co-director, Center for Microbiome Informatics & Therapeutics
Institute Member, Broad Institute of Harvard and MIT

Accepted by...
Christopher Burge
Professor of Biology and Biological Engineering
Director, Computational and Systems Biology Graduate Program

Linking Sequence to Function in Microbial Genomics

by

Sarah Jean Spencer

Submitted to the Program of Computational and Systems Biology
on July 28, 2017 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computational and Systems Biology

Abstract

Microbial genomes show high plasticity due to horizontal transfer, large community sizes, and rapid growth paired with adaptive mutations. Despite this mutability of gene content, most studies of microbial communities still rely on bulk, single-gene amplicon sequencing. In this thesis, I present methods that interrogate the gene content of single cells derived from complex natural communities. In the first project, I present a novel molecular biology method to link a bacterial functional gene to its host species with single-cell resolution. This high-throughput protocol is applied to assess the distribution of anaerobic respiration genes in a lake ecosystem. In the second project, I demonstrate extensions of this methodology to link genes between spatially proximal microbial cells, and apply this approach to probe the spatial organization of human dental plaque using DNA sequencing. In the final project, I completed whole-genome sequencing of environmental isolates derived from single, cultivable cells and employ mutational and horizontal transfer analysis to demonstrate adaptation to harsh environmental conditions in contaminated groundwater. These projects demonstrate the rich information stored within each microbial genome and the impact of spatial distribution in the environment. Each effort also contributes or highlights new molecular biology techniques to generate genomic data from individual microbial cells.

Thesis Supervisor: Eric Alm

Title: Professor of Biological Engineering

Acknowledgments

First I want to thank my mentors, past and present, who believed in my work and gave me the courage to keep improving and reaching father ahead. When I started out at Washington University in St. Louis, I'm not sure that I would have continued into graduate school without the encouragement of Prof. Sarah Elgin and Prof. Michael Brent. Prof. Elgin was the first academic to take a personal interest in my career and gave me the tough love and timely advice I needed to set my path towards graduate study. Prof. Brent was my first true advisor in that he gave me immense freedom to learn what I wanted, collaborate broadly, and jet set for interviews along with a strong recommendation – I'm positive my time in the Brent Lab boosted me into the CSB program. When I did arrive for graduate school, I was lucky enough to land into a group of incredibly supportive postdoc mentors. Prof. Ilana Brito gave me enough life advice for a lifetime, and Prof. Sarah Preheim was always around to correct, improve, teach, and shape my early work into mature research.

More than a mentor, more than a collaborator, I can't send enough thanks to my colleague and friend Manu Tamminen. Neither of us chose an easy path in research, but the energy, creativity, and excitement Manu brought to his work couldn't help but draw me in. I'm so grateful for the early training in lab, for the hours of wondering at the chalk board, for introducing me to colleagues across Boston. Without Manu's generosity and collaborative spirit, I would have had a much harder journey through grad school. Since he moved away, Manu has continued giving me thumbs up on skype after skype even when experiments weren't working. We made a great team; me with my skepticism and methodical nature, and Manu with his creativity and positive outlook. It's going to be hard to move on to less fun collaborations, but I'm so grateful to have such a good friend moving forward.

Over all these foundations, I'd like to thank Prof. Eric Alm for giving me the chance to train under him. In graduate school, I hoped to become more independent and creative, and I couldn't think of a better training ground than Eric's lab. He supported me on so many occasions, boosting me up when I hit bottom, funding my work despite high risks, giving me opportunities that far exceeded what graduate students typically get exposed to, and even flying me around the country and the globe to share my research. He didn't put undue pressure when I spent months facing challenges in lab, but always called to cheer me on the minute I had success. From the beginning, Eric met my work with a powerful mix of toughness and humor that make me feel immensely prepared for the next stage.

Eric's spirit and priorities also led him to assemble a group of people in lab that I can't imagine graduate school without. I overlapped the most with Sean Kearney and Mariana Matus, who were always willing to drop everything to give advice or trouble-shoot a problem with me. Thanks to Mathilde Poyet and Tu Nguyen for being incredible at cheering me up and supporting my work. Thanks to Scott Olesen for getting us both laughing out loud every time our paths cross. Many, many thanks to Jay Zhao and Tami Lieberman for stepping in at the last minute to help make my final chapter a success. Thanks Claire Duvallet for your positivity, passion, and GIFs – you're wonderful, don't change. And extra acknowledgments to Thomas Gurry, who gently addressed my enumerable computational issues as I learned the ropes, along with literally everyone else in lab. Finally, thanks to those administrators and managers who helped make all this possible, especially Shandrina Burns and Astrid Terry.

I want to thank my committee and my graduate program for their insight and support over the years. Prof. Chris Burge and Jacqueline Carota have been fantastic resources and shepherded me through especially the beginning and end of this experience. Thanks to Prof. Manolis Kellis for serving during my qualification exams and to Prof. Tim Lu for bringing so much energy and a fresh perspective to every meeting. I'm especially grateful to those who oversaw my thesis and defense, including Prof. Otto Cordero and Prof. Ilana Brito. Finally, my committee chair Prof. Paul Blainey was the best chair I could have asked for, and gave so many positive and helpful suggestions to lift up and advance both my research and my career.

Beyond the lab, I'd like to mention my personal rocks and foundations, my closest family and friends during this time. My family, Mom, Dad, Matt, and Annie, have been with me every step of this journey. I'm so grateful for Mom and Dad flying out every few months to shower me with positive support, and for all those wonderful speakerphone calls to check in and make sure everything was right. Thanks Matt for raging and praising and echoing all the intense emotions of grad school with me, and thanks Annie for answering my late-night calls and building me back up every time we talked. I want to also thank my aunt Barbara Beatty for believing in me enough to nominate me for additional funding – I value your time and energy so much. And of course, thanks to what I consider my closest family, my partner Michael for your patience, kindness, and unwavering support every single day, through the good and especially the bad days; my graduate experience, my career, and my life are so much better with you.

Last but not the least in any sense, I want to end by lifting up Ali Perrotta. A big reason I joined the lab was because of Ali – I could tell she was someone that would make the environment around her better no matter what. Ali and I shared so much over the years (five years!), going all the way from what we ate for breakfast to how we were dealing with choices in our lives and our careers. It's hard to imagine all those years of walking into the office without picturing some sunshine and energy from Ali's appearance. Not only that, but she helped me solve problems with my experiments, my collaborations, and basically every aspect of the complicated graduate experience. Spending time with Ali made work feel less like work, and made every day feel more like living instead of just surviving. I'm so grateful for her friendship, and I'm looking forward to being firm supports for each other no matter where we end up down the line.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Microbial genomes are characterized by substantial functional and spatial plasticity

Microbial functional diversity is enormous and influences environments from crevices in the human body to the global oceans. Even within a relatively constant, confined environment such as the human gut, estimates place the number of bacterial genes at 9 million, over 400 times the total number of human genes (1). These numbers increase exponentially when considering the diversity of environments available to bacterial communities globally, with modern estimates claiming as many as 1 trillion unique bacterial species on earth harboring an even larger functional gene pool (2). The study of these species remains an open area of discovery due to the vastness of species and chemistries, along with the complexity of assembled communities, with new phyla and functions being discovered every year (3,4).

Functional genes can change within and between hosts rapidly using mechanisms including mutation, horizontal transfer, and phase variation. Mutational processes can quickly alter the functional or regulatory capacity of microbes, whether through background mutation rates that sweep due to environmental pressure, or hypermutation that allows bacteria to explore a broader fitness landscape (5–7). Horizontal transfer is another dominant means of exchanging genetic material in real time, and separates core from flexible genomes in microbial species (8). While these are the most well-known mechanisms of functional plasticity, there are other means of rearranging functional genes internally via recombinases (9). Any one of these events can be difficult to detect, especially if occurring in rare members of a community, but these alterations carry the recent history of adaptation in perturbed environments.

The spatial structuring of bacteria within their local communities can also regulate access to functional gene content. On macroscales and across environmental gradients, microbes gain and lose functional capacity based on the changing environmental pressures. Oxygen gradients are one example, where nitrate reduction and other electron acceptor pathways increase as oxygen levels decrease (10). On physical scales relevant to individual bacterial cells, spatial structuring can influence the distribution of metabolic genes because closely associated cells can rely on diffusion and transport to share resources and intermediate products in pathways (11). Examples include the shared acquisition of iron with siderophores (12), and rampant auxotrophic relationships in multi-species biofilms (13,14). Both macroscale gradients and microscale structuring can shape bacterial associations, such as in human or environmental biofilms where an oxygen gradient forms across layers only a few cells deep and drives reproducible cross-phylum formations (15,16).

The study of bacterial spatial distributions and functional gene content is challenging due to enduring technical limitations, leaving many opportunities for molecular biology method development. Cell sorting and amplification allow users to link target functional genes with their hosts, but remain restricted in throughput. Innovations in fluorescent microscopy are beginning to reveal biological spatial structures, but can only capture a tiny fraction of characterized microbial species at a given time. The gold standard for functional content remains bacterial whole genome sequencing, however this generally requires cultivable target species and incurs high library prep and sequencing costs. In this thesis, I present three studies of bacterial functional and spatial structure that highlight innovative new molecular biology techniques and methods to increase throughput by reducing reaction volume. The following chapters each contain a unique application in environmental or human microbial communities that show the critical need to link genes to hosts, and link hosts to local community members and microenvironments.

1.1 New emulsion techniques link functional genes to host species in high throughput

In the second chapter of this thesis, I present a novel molecular biology technique designed to link a target functional gene to its host microbial species. The majority of recent sequencing efforts have focused on cataloguing species and bulk community composition, but functional inference has been difficult to reliably incorporate. Many functional genes are not perfectly preserved on a vertical phylogeny (17), and others are characterized by high rates of exchange (18,19). While methods exist to sequence whole genomes from single cells or cultured isolates in order to generate functional gene profiles (5,20), they limit throughput and generate megabases of sequence data that may be irrelevant to a targeted research question.

As an alternative, I developed epicPCR (Emulsion, Paired Isolation, and Concatenation PCR), which is an emulsion-based method to physically concatenate two target genes from a single genome. This method relies on high dilution rates into millions of emulsion droplets to achieve single-cell isolation, then employs an acrylamide encapsulation step to allow a variety of microbial lysis techniques. Encapsulated, lysed genomes are resuspended in a PCR emulsion with three primers designed to amplify and stitch together two target genes. The specificity of this protocol was tested with synthetic amplicon beads, and then we applied to protocol to identify bacterial species carrying a dissimilatory sulfite reductase gene in the anoxic region of a stratified lake. The protocol design is versatile and allows different primers as well as cell handling and loading to test a variety of hypotheses.

1.2 Adaptation of emulsion methods towards spatial sequencing of biofilm aggregates

The third chapter of this thesis presents work building towards a sequence-based readout of bacterial cell-cell associations at the microscale. Virtually all sequence-based assays in the microbiology community target bulk collections, single cells, or cultivated isolates. There is a missing component of information in the way individual cells aggregate to cooperate or compete in their local microenvironment. Spatial assembly at the micron-scale can serve a variety of functions including providing protection from antibiotics, enabling horizontal transfer, and supporting close cross-feeding relationships (21–23). However current techniques to provide this community structure information fall short, relying on limited genus-level microscopy or low throughput aggregate sorting (24,25). Even with the low resolution information from these techniques, valuable insight about community dynamics has been revealed, such as rich CO_2, lactate, and acetate exchange in the microaerophilic perimeter of human oral plaques (26).

To test a new approach enabling high-throughput, microscale spatial sequencing, I adapted the epicPCR protocol to record cell-cell associations in suspended microbial aggregates. First, I present a clade-targeted primer design analogous to the original epicPCR design, in which an initial target gene in a restricted clade is fused to a universal 16S rRNA gene fragment. This design was replicated to target both a well-characterized clade, namely the *Streptococcus* genus, and a candidate phylum, *TM7*. In both cases, a restricted and partially replicated set of associated bacterial species was recovered, which differed based on the sampling site. Next I present results from two barcoding designs, in which a droplet-specific barcode is amplified and linked to any available 16S rRNA gene in the same droplet. These studies, when corrected for any high bulk species abundance, show specific and partially reproducible patterns of bacterial co-localization, and will inform future efforts to generate this novel data type.

1.3 Whole genome sequencing of environmental isolates reveals recent adaptive changes *in situ*

The fourth chapter of this thesis reveals the power of low reaction volume, high-throughput whole genome sequencing in discovering recent adaptation along an environmental gradient. In order to move toward total genomic awareness of *in situ* communities, it's critical to produce whole genome sequences of active isolates from the site. These data provide a snapshot of what a single, viable cell was functionally capable of at the time of collection, including the flexible genome and any acquired plasmids. Whole genome sequences also allow powerful comparative genomics between closely related isolates, enabling the identification of strains, adaptive mutations, and species migration patterns that occurred on recent evolutionary timescales across a macroscale sampling area (5,27). Overall, whole genome sequencing can highlight the immense functional plasticity and metabolic richness carried by individual and closely related isolates.

In this chapter I generated draft genomes for 265 isolates cultivated from groundwater wells spanning a nitrate and heavy metal gradient. Of the isolates, 139 grouped into the *Pseudomonas* genus and populated fifteen strain-level subgroups. Within each strain, I identified high confidence single nucleotide variants (SNPs) between the isolates and found groupings of non-synonymous mutations that were reproduced or fell along a common pathway. This revealed a strong signal for regulatory alterations affecting iron acquisition in one well. I also identified instances of gene loss differentiating isolates from the same strain, which also impacted transcriptional regulation. Each of these adaptive changes can be linked to both a physical sampling site and a phylogenetic strain, enabling functional and dispersive discovery across the landscape of the sampling region.

1.4 Assays of microbial function at scales relevant to individual cells move functional discovery beyond inference

In a field dominated by big data from bulk communities, this thesis explores new and novel molecular techniques to add throughput, resolution, and functional structure to microbial genomics. These and similar techniques have generated considerable interest in the research community, with more efforts to miniaturize and target single-cells as well as micro-aggregates. The versatility of these molecular protocols means that there are many exciting future directions for this work, including novel primer designs, new hydrogel formulations, and opportunities for automation. In the final chapter of this thesis, I discuss the remaining challenges as well as future implications of miniaturized functional genomics, and suggest further efforts to heighten impact, reduce cost, and disseminate these techniques to the broader microbiology community.

Chapter 2 Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers

Sarah J Spencer, Manu V Tamminen, Sarah P Preheim, Mira T Guo, Adrian W Briggs, Ilana L Brito, David A Weitz, Leena K Pitkänen, Francois Vigneault, Marko P Virta, Eric J Alm

The work presented in this chapter is available as a manuscript published in the *ISME Journal* (2016) volume 10, pg. 427-436.

Abstract

Many microbial communities are characterized by high genetic diversity. 16S ribosomal RNA sequencing can determine community members, and metagenomics can determine the functional diversity, but resolving the functional role of individual cells in high-throughput remains an unsolved challenge. Here, we describe epicPCR (Emulsion, Paired Isolation, and Concatenation PCR), a new technique that links functional genes and phylogenetic markers in uncultured single cells, providing a throughput of hundreds of thousands of cells with costs comparable to one genomic library preparation. We demonstrate the utility of our technique in a natural environment by profiling a sulfate-reducing community in a freshwater lake, revealing both known sulfate reducers and discovering new putative sulfate reducers. Our method is adaptable to any conserved genetic trait and translates genetic associations from diverse microbial samples into a sequencing library that answers targeted ecological questions. Potential applications include identifying functional community members, tracing horizontal gene transfer networks, and mapping ecological interactions between microbial cells.

2.1 Introduction

"Who is doing what" is a major open question in microbial ecology. While 16S ribosomal RNA sequencing can answer the "who", and shotgun metagenomics can partially address the "what", connecting the two is difficult. In recent years, investigators have tried different approaches to ask targeted ecological questions at the resolution of single cells. The most common approach to connect phylogeny with function combines single cell FACS sorting with whole genome amplification and PCR screening for target genes (28–31). Other methods isolate single cells using microfluidics, then screen for target genes either in microfluidic chambers or on primer-coated beads (32–34). There are also variants of fluorescence in situ hybridization (FISH) that show co-localization of target gene probes (35–37). Despite these advances, current methods face persistent limitations in throughput, reagent costs, and labor requirements. Motivated by this technology gap, we developed a cost-effective and highly parallel technology to answer "who is doing what" in high-throughput in any microbial community.

Here we present epicPCR (Emulsion, Paired Isolation, and Concatenation PCR), a novel method for recovering linked phylogenetic and functional information from millions of cells in a single experiment. Emulsion-based techniques provide a simple way to partition bulk reactions into millions of individual reactions, each within a single droplet. This approach is not new, and has been used by sequencing platforms such as 454 and Ion Torrent to prepare templates for sequencing. Emulsion techniques have also been employed in studies of human haplotypes from single cells and studies of single-cell immunology (38,39) using emulsions in combination with fusion PCR, a technique originally developed for preparing fusion proteins (40).

A significant challenge in translating emulsion technology to microbiology is the difficulty of microbial cell lysis. The epicPCR methodology we present here permits efficient cell lysis by isolating cells in emulsion droplets prior to PCR and encapsulating them in a hydrogel matrix (41). This matrix is dense enough to hold bacterial genomes in place after hydrogel bead recovery, but loose enough to allow enzymes and primers to diffuse through (42,43). Hydrogel beads are then loaded into a second emulsion where amplified target genes become physically linked by fusion PCR.

We demonstrate epicPCR by detecting a rare sulfate reducing cell population among the microbial diversity of a freshwater lake, sequencing 16S ribosomal RNA (rRNA) genes from cells containing the dissimilatory sulfate reductase gene *dsrB* (44). We confirm that the observed phylogenetic distribution of *dsrB* genes matches predictions based on observed geochemistry, while also revealing previously undetected putative sulfate reducers. The efficiency of microbial cell lysis can be measured by comparing untargeted epicPCR with bulk 16S rRNA gene data. Our bulk emulsion design can query hundreds of thousands of cells in parallel with costs comparable to one genomic library prep, increasing throughput and reducing expense compared to existing methods. This adaptable method can translate genetic associations from any sample into a sequencing library that answers targeted ecological questions.

2.2 Results

2.2.1 Benchtop emulsions enable genome capture and targeted sequencing of single cells within complex communities

epicPCR combines established methods for cell isolation, encapsulation, and paired amplification. An overview of the method is as follows: An initial aqueous sample-in-oil emulsion generates approximately 500 million droplets, each about one nanoliter in volume, that contain single cells. These cells are loaded and dispersed assuming Poisson statistics, so that on average less than one droplet in 100 contains a cell. Each of these droplets also contains acrylamide monomers which polymerize and encapsulate cells upon addition of a catalyst, forming polyacrylamide beads (Fig. 2-1A). The polyacrylamide hydrogel provides support for bacterial chromosomes and plasmids, preventing their diffusion when the trapped cells are combined in bulk and redistributed for fusion PCR. The diameter of the polyacrylamide beads typically ranges from 5 to 30 μm with most beads having a diameter around 10 μm, determined by light microscopy as previously described (41), with representative images in Fig. A-4A.

Fusion PCR is performed on the hydrogel-trapped genomes in a secondary emulsion (Fig. 2-1B, Fig. A-4B) to ensure that each epicPCR is compartmentalized (Fig. 2-1C, D). The protocol has been described previously (38) and proceeds as a single reaction with an initial linear amplification of the 16S ribosomal RNA gene and a limited-cycle exponential amplification of a separate target gene. The limited-cycle exponential amplification is done using a primer pair where one of the primers has an overhang that is complementary with a part of the 16S rRNA gene. After this overhang-primer is depleted, the complementary part will form a fusion amplicon with the 16S rRNA gene, and exponential amplification of the fusion amplicon proceeds.

Illumina adapters are subsequently added to pooled fusion amplicons in a bulk nested PCR (Fig. 2-1E; Fig. A-1B). Without refined molecular control, partially fused products could continue the reaction in bulk and destroy single-cell specificity. Aptamer-based hot start polymerase prevents partially fused products from extending, preserving single-cell specificity in the bulk reaction. Then a saturating concentration of blocking primers anneals to and removes any partially fused pieces from the bulk library amplification (39,45) (Fig. A-1B). Collectively, the steps of this protocol are designed to preserve the individually fused information from single cells while maintaining high throughput.

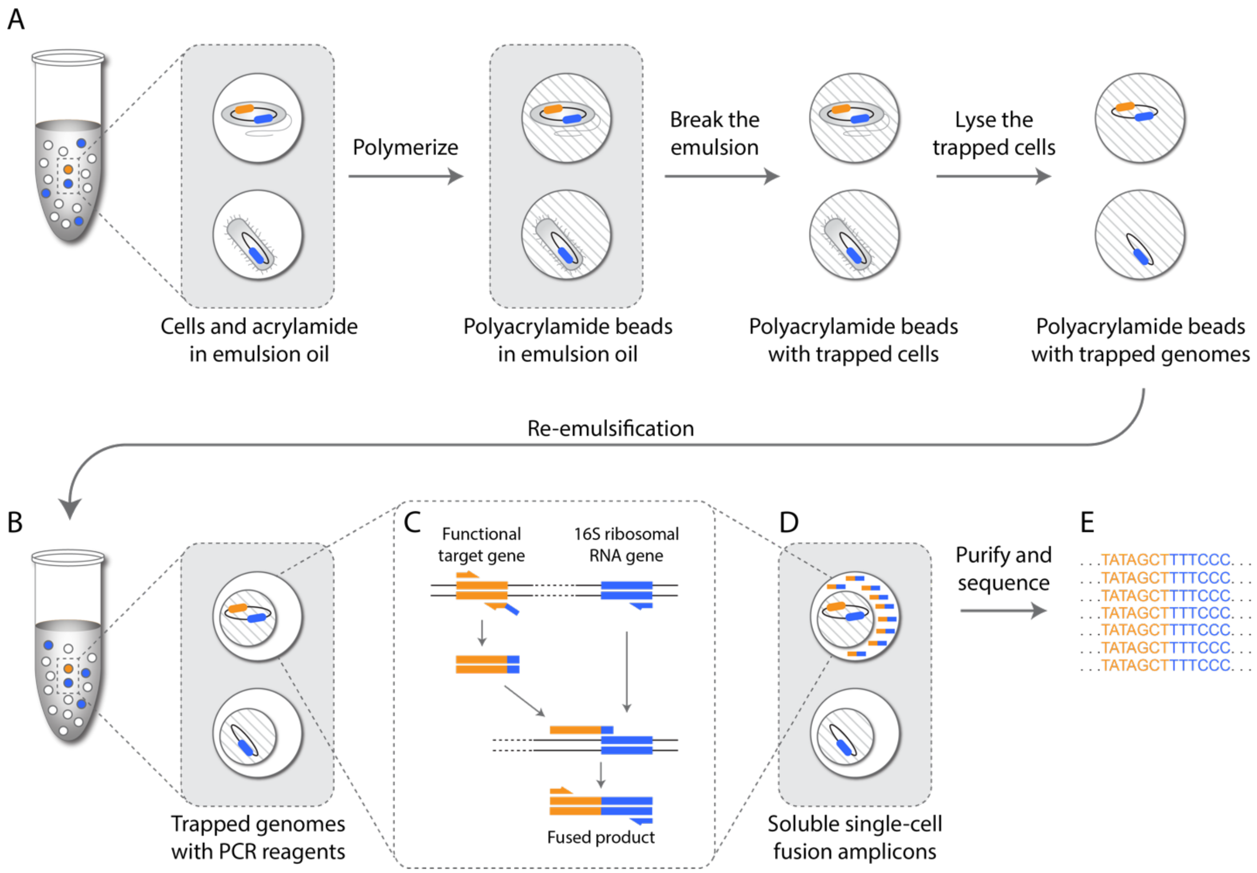

Figure 2-1. Workflow of epicPCR. A) Microbial cells in acrylamide suspension are mixed into emulsion oil. The emulsion droplets are polymerized into polyacrylamide beads containing single cells. The emulsion is broken and the cells in the polyacrylamide beads are treated enzymatically to destroy cell walls, membranes and protein components, and expose genomic DNA. B) Polyacrylamide-trapped, permeabilized microbial cells are encapsulated into an emulsion with fusion PCR reagents. C) Fusion PCR first amplifies a target gene with an overhang of 16S rRNA gene homology. With a limiting concentration of overhang primer, the target gene amplicon will anneal and extend into the 16S rRNA gene, forming a fusion product that continues to amplify from a reverse 16S rRNA gene primer. D) The fused amplicons only form in the emulsion compartments where a given microbial cell has the target functional gene. E) After breaking the emulsion the fused amplicons are prepared for next-gen sequencing. The resulting DNA sequences are concatemers of the target functional gene and the 16S rRNA gene of the same cell.

2.2.2 Spiking an environmental sample with synthetic control beads demonstrates high specificity of epicPCR

One exciting application of this technology is to link phylotype to function in a complex community. Here we processed lake water from oxic and anoxic depths, then used epicPCR to target cells harboring the dissimilatory sulfite reductase gene *dsrB*. Sulfate reduction is a process where microbial cells in anoxic conditions use sulfate as the terminal electron acceptor of their metabolism. We recorded the geochemistry of water from an urban lake by measuring sulfate, nitrate and oxygen at one-meter intervals down to 22 meters (see Section A.1.6 for details). At a 21 m depth, both oxygen and nitrate are depleted, but sulfate is still available as an electron acceptor (Fig. A-5).

Our single-cell experimental design consisted of epicPCR assays on 2 m and 21 m lake water with positive and negative spike-in controls. We produced spike-in controls by synthesizing polyacrylamide beads that contained covalently attached DNA amplicons. Negative control beads carried a mock-16S rRNA gene whereas positive control beads had both a mock-16S rRNA gene (with a sequence distinct from the negative control beads) and a mock-*dsrB* sequence.

To compare the full 16S rRNA gene diversity present to the *dsrB*-carrying subpopulation, we completed both non-specific and *dsrB*-specific epicPCR assays. Our non-specific assay fused together 16S rRNA gene sequences with a synthetic amplicon carrying a random DNA barcode. The barcode, based on 20 degenerate nucleotides, was added at a concentration of 10 pM, which loads on average three molecules per 10 μm diameter droplet. Since cell-containing and control polyacrylamide beads are all likely to be in droplets containing barcodes, we expected this reaction to result in fusions to all environmental, positive and negative control 16S rRNA gene sequences. Our *dsrB*-specific assay fused *dsrB* gene fragments with 16S rRNA genes present in the same droplet. We expected to observe only 21 m, anoxic species and positive control 16S rRNA gene sequences in our *dsrB*-fusion products.

Fusions to 16S rRNA genes from environmental cells matched our expectation that sulfate-reduction machinery would only occur at anoxic depths. We recovered *dsrB*-16S fusion amplicons from the 21 m depth, but detected no *dsrB*-16S rRNA gene fusions (abbreviated *dsrB*-16S) at 2 m (Fig. 2-2). The depth specificity is not due to assay bias because 1 167 006 non-specific barcode-16S fusions evenly captured both 2 m and 21 m diversity.

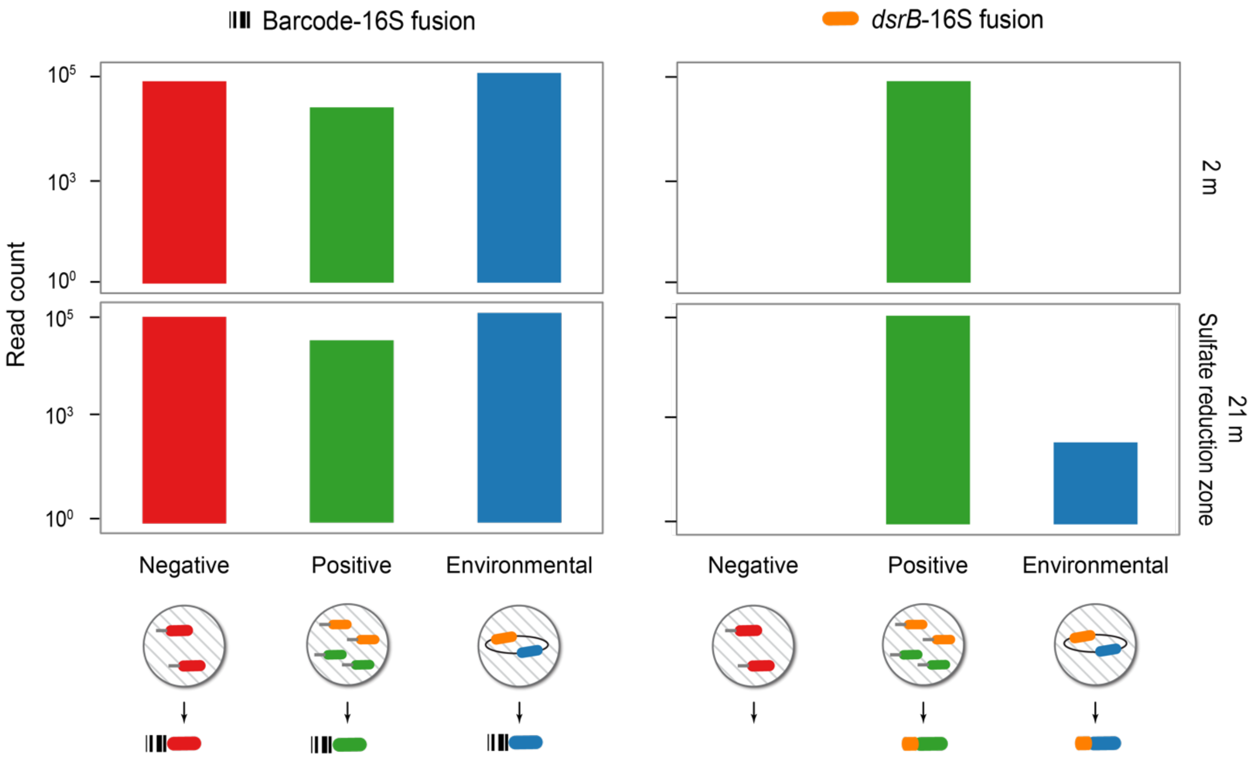

Figure 2-2. Specificity of epicPCR is tested in a series of experiments in which a random barcode or a *dsrB* gene fragment is fused with the 16S ribosomal RNA gene in an environmental sample that is spiked with negative and positive controls. Negative controls are synthetic polyacrylamide beads with attached mock-16S amplicons. In epicPCR these beads result in a positive signal for barcode fusion but give no signal for *dsrB*-16S fusion. Positive controls are synthetic polyacrylamide beads with attached mock-16S and mock-*dsrB* amplicons. In epicPCR these beads result in a positive signal for both barcode-16S and *dsrB*-16S fusions. For environmental cells from a freshwater lake, barcode-16S reactions capture the 16S rRNA gene diversity at both 2 m and 21 m depths. Sulfate reduction takes place in the anoxic layers far below the surface, so *dsrB*-16S fusions only occur successfully at the 21 m depth.

As expected in our controls, we observed ubiquitous 16S rRNA gene fusions to the non-specific barcode amplicon, but highly specific positive control amplification in *dsrB*-16S fusion products. Barcode-16S fusion products captured 388 768 reads containing the negative control 16S rRNA gene sequence and 70 154 reads containing the positive control 16S rRNA gene sequence. In contrast, the targeted *dsrB*-16S fusion design captured exclusively positive control 16S rRNA gene sequences – a total of 372 223 reads – with zero observations of the negative control 16S rRNA gene sequence, confirming the high specificity of the technique.

2.2.3 Abundant phyla are consistently targeted by epicPCR

Comparisons of the 16S rRNA gene diversity from barcode fusion and bulk 16S rRNA gene sequencing shows that epicPCR recovers all major phylogenetic groups, indicating that cells from most of these groups became successfully permeabilized in a replicated experimental setup despite variable cell wall structures (Fig. 2-3, Fig. A-6). Treatment with lysozyme, proteinase K, detergents, and heat permeabilized certain additional phyla relative to the standard epicPCR protocol.

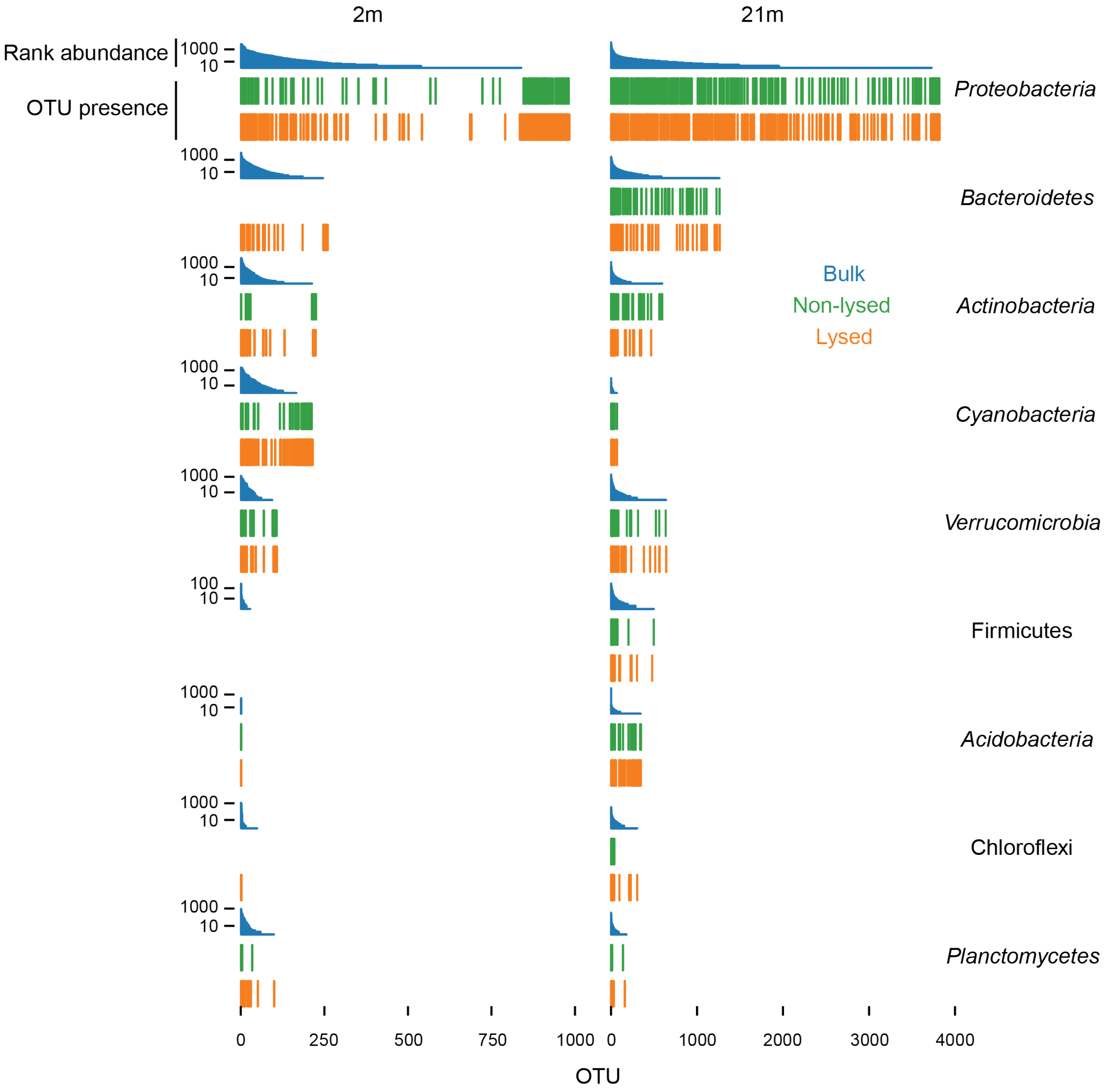

Figure 2-3. Bacterial groups recovered by a bulk 16S rRNA gene survey and epicPCR from the 2 m and 21 m depths. OTU rank abundance of the bulk 16S rRNA sequencing is presented as blue histograms. Corresponding OTUs identified by epicPCR are presented as bars below the rank abundance histograms. This includes reactions with (yellow) and without (green) additional lysis reagents. epicPCR captures most phyla within a sample, regardless of cell structure or phylogeny. The use of additional lysis reagents including lysozyme, proteinase K, and detergents, increases the

phylogenetic coverage of the assay for certain bacterial groups such as *Actinobacteria*, *Bacteroidetes*, *Chloroflexi*, *Cyanobacteria* and *Planctomycetes*.

Most dominant phyla were successfully permeabilized even without enzymatic treatment (Fig. 2-3). However, certain phyla such as *Actinobacteria*, *Bacteroidetes, Cyanobacteria* and *Planctomycetes* at the 2 m depth and *Chloroflexi* at both depths required additional enzymatic lysis for improved operational taxonomic unit (OTU) recovery. We also note that *Firmicutes* at 2 meters produced no reads regardless of permeabilization. Due to low OTU recovery with bulk sequencing of this group, we suspect this was a result of sampling bias rather than actual resistance of this phylum to epicPCR. We hypothesize that the *Proteobacterial* and *Cyanobacterial* OTUs at 2 meters that were present in epicPCR experiments but not in bulk 16S sequencing result from the lower coverage of the bulk 16S sequencing.

Polyacrylamide formation and thermal cycling with additional enzymatic lysis proved sufficient to reproducibly recover rare candidate phyla, including H-178 with a 16S rRNA gene bulk read abundance of 7.8×10^{-4} (data not shown). epicPCR recovered this rare taxon using the non-specific, barcode-16S assay design. Thus the targeted, functional fusion approach could selectively amplify rare phyla and species to a much greater proportion of the final sequence data.

2.2.4 epicPCR links metabolic functions to known and putative hosts

We repeated the *dsrB*-16S fusion on a larger number of cells to profile the lake water sulfate reducing community. To confirm that epicPCR targets a wide range of bacterial reducing *dsrB* genes, we tested the primers *in silico* to a database of known *dsrAB* genes (46) and compared the epicPCR *dsrB*s to bulk *dsrB* sequences (Fig. A-3B). *In silico* PCR confirms that epicPCR primers have a broad specificity across bacterial reductive *dsrB*s but do not amplify bacterial oxidative *dsrB*s or archaeal reductive *dsrB*s. We observe an overlap between the bulk and epicPCR *dsrB* sequences and conclude that epicPCR targets a wide variety of reductive *dsrB* sequences in the lake water belonging to the *Deltaproteobacterial dsrB* supercluster (Fig. A-3B). We suspect that the few hits of epicPCR *dsrB*s to oxidative or archaeal *dsrB*s result from low phylogenetic information of the *dsrB* fragment rather than low specificity of the epicPCR primers.

From the same set of *dsrB*-16S fusion sequences, we analyzed the 16S rRNA genes to test whether our observations include known sulfate-reducing bacteria. A maximum likelihood analysis (FastTree2 (47)) grouped the epicPCR 16S rRNA gene sequences within the *Deltaproteobacterial* families *Syntrophobacteraceae, Syntrophaceae and Desulfobacteraceae* (Fig. 2-4), members of which have been confirmed to contain the *dsrB* gene (48). Phylogenetic analysis against a database of known sulfate reducing bacteria (46) revealed that 319 364 out of 2 028 199 sequenced amplicons have less than 95% similarity to the closest known sulfate reducer and thus represent novel OTUs. Both novel and non-novel OTUs primarily have their closest matches in the Greengenes database to *Deltaproteobacteria* (Table A-6), indicating that novel groups found by epicPCR are likely not false positives. We also detect a fraction of 0.2% of *Gammaproteobacterial* and *Betaproteobacterial* reads that are most likely an unspecific background of the method.

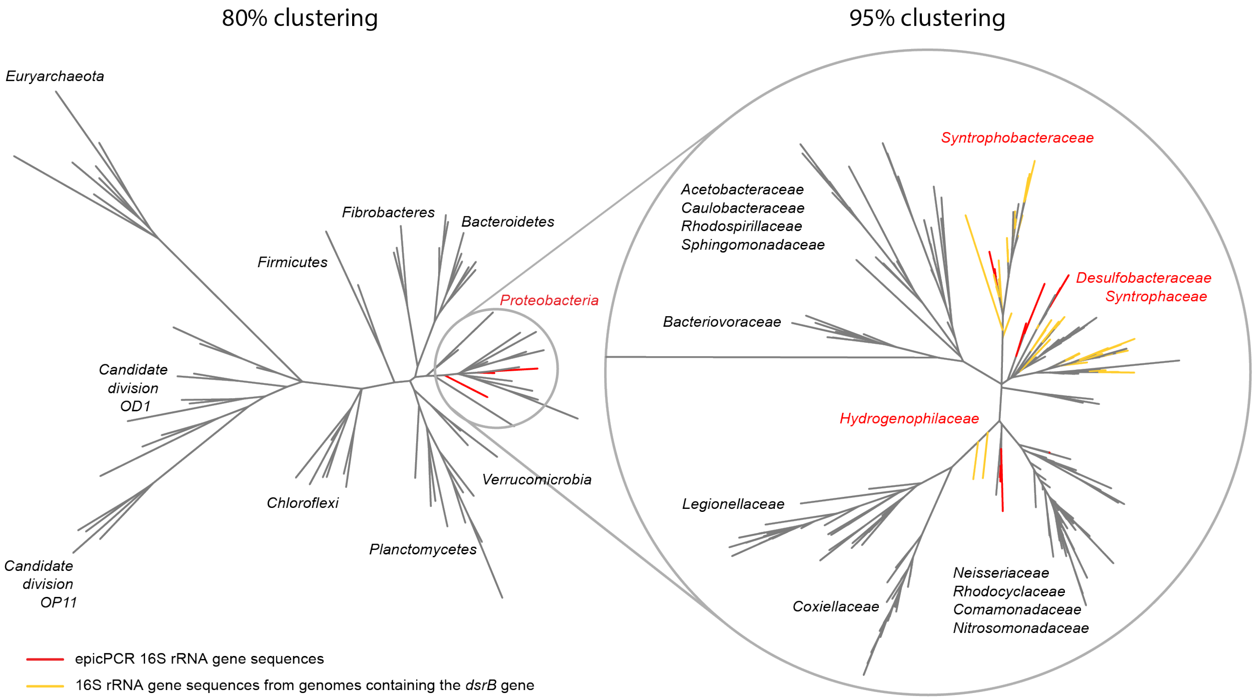

Figure 2-4. A maximum likelihood tree of the microbial diversity in lake bottom water (21 meters). The tree was constructed from the total 16S rRNA gene sequences from lake bottom water clustered

by 80% and 95% similarity, 16S rRNA gene sequences belonging to known sulfate reducing species (yellow branches), and 16S rRNA gene sequences recovered by epicPCR by the presence of *dsrB* (red branches). The 16S rRNA gene sequences recovered by epicPCR group within *Proteobacteria* with members from families *Desulfobacteraceae, Syntrophaceae, Syntrophobacteraceae*, that have previously been confirmed to contain the reductive *dsrB* gene (48,49).

2.3 Discussion

Keeping pace with sequencing improvements, 16S rRNA gene surveys and metagenomic surveys are now being enriched with methods to separate and characterize the function of single cells within complex populations. Here we describe epicPCR, a novel technique to connect microbial function to phylogeny in a simple, high-throughput protocol. Using the highly parallel nature of emulsions, epicPCR provides a throughput of millions of cells with the cost of a single sequencing library preparation. We confirm the high specificity of epicPCR using synthetic control beads, then successfully enrich for a collection of sulfate-reducing prokaryotes in the anoxic region of a stratified lake.

Key technological advances that are critical for an optimal performance of epicPCR include hydrogel formation and re-emulsification for fusion PCR, and certain optimizations for bulk downstream amplification. Sufficient dilution of cells or hydrogel beads prevents emulsion overloading, and adding glass beads into the tube during secondary emulsion production provides additional shear force to separate hydrogel beads into individual droplets (see Section A.1.4). A three-primer fusion design ensures that only droplets containing a target gene produce amplicons, reducing unwanted 16S rRNA gene artifacts in the bulk mixture. Blocking primers, with highly efficient 3' 3-carbon-spacer blocks, also inhibit the spurious, chimeric amplification of incomplete fusion products within bulk reactions (39,45).

epicPCR can determine the hosts of any target gene with conserved priming sites, and extensions of the method could generate quantitative or novel co-occurrence data. Our primer design only captured a small fraction of the *dsrB* gene, but an updated design could capture a long enough region of the target gene to construct dual target gene and 16S rRNA gene phylogenies in order to demonstrate coevolution or ancient horizontal transfers. Due to the non-linear effects of amplification and droplet size, the generated data forms a qualitative list of species rather than quantitative ratios. We expect that controlling droplet size with microfluidic droplet makers or tagging droplet products with molecular barcodes could produce quantitative results.

A variety of ecological questions become accessible with epicPCR, including which species drive biogeochemical cycles, harbor integrated phage, or carry antibiotic resistance genes. While these topics would require dispersing cell aggregates into single cells (as described in (50) and (51)), we also envision adaptations of epicPCR that would target more than one genome. epicPCR could query for host associations such as microbe-protist interactions by fusing 16S and 18S ribosomal RNA genes. By fusing a random barcode with 16S rRNA genes when targeting cell aggregates, fused 16S rRNA gene sequences under a single barcode would indicate physical co-occurrence and therefore spatial structuring of bacteria.

More general physical co-occurrence data could be collected by concatenating targets beyond bacterial genomic DNA. Combining the epicPCR concept with cDNA synthesis, the technique could have applications in immunology, including assaying the co-occurrence of T-cell receptor variable regions and T-cell master regulators (39). Extending from our current protocol, attachment of different functional molecules such as PCR primers or antibodies to the hydrogel matrix could lead to completely novel experimental strategies.

2.4 Materials and methods

2.4.1 Lake water sample collection and quantification

Lake water was collected from Upper Mystic Lake (~ 42 26.155N, 71 08. 961W) near Winchester, MA on August 12, 2013. Duplicate samples were taken from 2 m and 21 m depths, with 15 ml of lake water immediately placed in 25% glycerol and frozen on dry ice for transport and subsequent storage at -80 °C. Approximate cell counts were determined using one of the duplicate samples for each depth. Samples were diluted, fixed with formalin, and stained with DAPI to perform cell counts on a fluorescent microscope. Description of DNA extraction and bulk 16S rRNA gene library preparation for these samples can be found in Section A.1.6.

2.4.2 Polymerization and lysis of lake water samples

We thawed a glycerol stock of lake water and suspended 14 million cells in nuclease-free water. This suspension was combined with ammonium persulfate, acrylamide, and N,N′ -Bis(acryloyl)cystamine as a crosslinker. The 255 µl aqueous mixture was applied to 600 µl Span 80/Tween 80/Triton X-100 emulsion oil (52) and vortexed for 30 s, which produced approximately 500 million droplets (based on 10 µm average droplet diameter, see Fig. A-4). We added a small volume of TEMED to catalyze the polymerization and vortexed for an additional 30 s, then let the emulsion polymerize for 90 min. Polyacrylamide beads were extracted with diethyl ether, then resuspended in 1 ml 1X TK buffer and filtered through a 35 µm cell strainer. Detailed methods for these steps are available in Section A.1.4.

We performed epicPCR assays on the polyacrylamide beads both with and without additional lysis reagents. For the beads with additional lysis treatment, we added 0.8% Ready-Lyse Lysozyme (35,000 U/µl, Epicentre, Madison, WI, USA) to polyacrylamide bead aliquots and incubated at 37 °C overnight. Each aliquot was centrifuged and resuspended in 1X TK buffer, then treated with 20% (v/v) proteinase K (1 mg/ml, Sigma, St. Louis, MO, USA) and 0.8% (v/v) Triton X-100. The samples were incubated at 37 °C for 30 min, then 95 °C for 10 min. Following treatment, polyacrylamide beads were again centrifuged and resuspended in 1X TK buffer for the epicPCR library preparation.

2.4.3 Preparation of synthetic control polyacrylamide beads

We amplified DNA segments with acrydited 5’ ends and attached them to polyacrylamide beads to serve as synthetic positive and negative controls. To prepare these beads, we created a bulk emulsion with approximately 500 million droplets by vortexing for a total of 60 seconds, and diluted our acrydited DNA to load 100 molecules per droplet on average. In our negative control preparation, we added 0.7 µM 16S-V4neg PCR product. In our positive control preparation, we added 0.7 µM 16S-V4pos PCR product plus 0.7 µM dsrB-synth primer (PCR product and primer sequences in

Table A-1, primers adapted from (53)). To prepare the polyacrylamide beads we combined our acrydited DNA segments with an aqueous reaction mixture, emulsified the aqueous phase, and polymerized the emulsion droplets as described in Section A.1.5. Five rounds of centrifugation (12,000 *g* for 1 min) and removal of the low molecular weight polyacrylamide beads, followed by filtration through a 35 μm cell strainer, ensured a more even size distribution for the synthetic controls.

2.4.4 epicPCR library preparation

First we prepared an emulsion with polyacrylamide beads and fusion PCR primers in order to amplify the single-cell fusion templates. The PCR mix included 45 μl of polyacrylamide beads combined with PCR reagents and emulsion stabilizers (BSA and Tween 20). We also added the three fusion primers (Fig. 2-1B; Fig. A-1A,D; Table A-2): 1 μM F1, 1 μM R2, and a limiting concentration of 10 nM R1-F2' to bridge between the target gene and 16S rRNA genes. These generic primer names refer to Fig. A-1A; for specific experiments, please refer to Fig. A-1C-D for primer names and Table A-2 for primer sequences. For PCRs with a soluble barcode-16S rRNA gene fusion (abbreviated barcode-16S), we added 100 fM fusionBarcode. Table A-3 presents an outline of primers used for different experiments and Figure A-1C-D shows fusion construct designs. Figure A-2 shows the genomic context of the *dsrB* primers, adapted from (54,55). The final aqueous PCR mix was added to 900 μl ABIL EM 90 emulsion oil (52), vortexed, and then aliquot into PCR tubes for thermocycling. Following amplification, aliquots were pooled, phase separated, and purified with AMPure XP beads (see Appendix A.1 for detailed procedures and sample information).

Following this reaction, we added another set of primers to nest within the fused products and also block the amplification of unfused pieces (Table A-4). The nested PCR included standard PCR reagents combined with 0.3 μM forward and reverse nested primers (for specific experiments, please refer to Fig. A-1C-D for primer names and Table A-4 for primer sequences) plus 3.2 μM each of U519F_block10 and U519R_block10, which are modified universal 16S rRNA gene primers (56) that prevent amplification of unfused pieces (Fig. A-1B). The blocking primers were enhanced from the design presented in (39,45) by the addition of 3' 3-carbon-spacers; these spacers show decreased degradation and increased blocking efficiency over 3' phosphates (57). We combined the nested and blocking primers with purified fusion product from the previous reaction and ran qPCRs to determine the number of amplification cycles to use for each sample. Using the qPCR Ct values, we completed the final nested reaction, purified the products, and amplified again with Illumina adapters (Table A-5). These adapters included a 3' YRYR sequence to add template diversity to the amplicon library. Purified final libraries were sequenced on an Illumina MiSeq with 250 bp paired-end reads (see Section A.1.4 for detailed procedures).

2.4.5 epicPCR sequence analysis and OTU clustering

Resulting sequence data was filtered for quality and expected fusion structure. Throughout analysis, we frequently used functions from the software package QIIME; functions had default parameters unless otherwise specified (58). After splitting samples by sample barcode, we stitched together forward and reverse reads and then filtered for quality (at Phred > Q20). Chimera checking was critical for our fusion constructs, so we ran the non-reference-based identify_chimeric_seqs.py (-m usearch61). Remaining reads were trimmed to 121 bp of the 16S rRNA gene V4 region based on a conserved 16S rRNA gene V4 site (59) and we discarded any reads that did not match our expected fusion bridge structure using custom python scripts (version 2.7; https://github.com/sjspence/epicPCR). In order to identify positive and negative control 16S rRNA gene sequences, we performed a targeted BLAST search against our synthetic 16S rRNA gene sequences.

For Operational Taxonomic Unit (OTU) determination, we first collapsed identical droplet barcode-16S pairs into a single representative sequence using a custom python script (version 2.7; https://github.com/sjspence/epicPCR/blob/master/compressBar.py). This function controlled for droplets that amplified exponentially more than others due to heterogeneous droplet volume. We then ran a series of QIIME functions that grouped 16S rRNA gene sequences into 97%, 95% and 80% identity clusters, picked representative sequences, and assigned taxonomy based on the Greengenes and SILVA databases (58,60,61). In order to facilitate visual comparison between samples despite different sequencing depths, we rarefied to the sample with the fewest reads; when we did not compare between samples, we presented the full read set (see Section A.1.5 for detailed procedures).

For tree construction the 16S rRNA gene sequences picked by epicPCR were combined with bulk 16S rRNA gene sequence data of the sample. The epicPCR 16S rRNA gene sequences and bulk 16S rRNA gene sequences had been separately grouped into 95% and 80% identity clusters, and sequences from the respective clustering distances were combined. The sequences were aligned using SINA (62). Tree construction was done using FastTree 2.1.7 (47).

For functional classification the *dsrB* sequences were grouped into 95% identity clusters by uclust 1.2.22 and aligned to a *dsrAB* database (46) using the NAST output option of usearch v8.0.1517 (63). A reference tree was constructed from the *dsrAB* database using FastTree 2.1.7 (47). Range and specificity of epicPCR primers (Fig. A-1D; Table A-2; Table A-4) was tested in an *in silico* PCR against the *dsrAB* database in two steps using the EMBOSS 6.5.7 primersearch tool with 20% mismatch cutoff (64): first we extracted *in silico* amplicons from the *dsrAB* database using the sequence of primer dsrB-F1 and segment 5'-TGCCTSAAYATGTGYGGYG-3' from primer dsrB-R1. Subsequently we extracted a subset from these *in silico* amplicons using primer segments 5'-VAGVATSGCGATRTCGGA-3' from i_dsrB-F3 and 5'-TGCCTSAAYATGTGYGGYG-3' from dsrB-R1. Complete matches of epicPCR *dsrB* fragments to the *dsrAB* database were identified using the grep tool from OS X Yosemite. Matches of bulk *dsrB* fragments (see Section A.1.6 and Fig. A-3A for details of bulk *dsrB* sequencing) to the *dsrAB* database were identified using BLAST 2.2.30 (65) using a similarity cutoff of 70%. The *in silico* PCR results, epicPCR *dsrB* matches, and bulk *dsrB* matches in the *dsrAB* database were visualized in a *dsrAB* reference tree using iTOL (66) (Fig. A-3B).

2.4.6 Data Access

The raw sequencing data from this study were submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA264605. The computational steps we used to process the data are detailed in a text file along with custom scripts available at https://github.com/sjspence/epicPCR.

Chapter 3 Towards spatial proximity sequencing in the oral microbiome

Sarah J Spencer, Floyd E Dewhirst, Manu V Tamminen, Eric J Alm

Abstract

The spatial proximity of bacteria, particularly within biofilms, informs their individual and community functional profile. Co-localized cells can engage in cross-feeding, parasitic, or competitive relationships, and may also assemble into protective microstructures to shield internal cells from external environmental effects. The human oral microbiome has traditionally remained on the forefront of biofilm research due to its ease of access and moderate complexity of microbial interaction, but current techniques for investigating spatial structure fall short of species-specific, high-throughput readouts. In this work we tested the ability to port emulsion-based encapsulation and molecular fusion techniques into a spatial readout of biofilm aggregates within the oral microbiome. First, we tested a clade-targeted design to recover co-aggregators with a known taxon, and generated putative connections to the *Streptococcus* genus and the parasitic *TM7* phylum. In a second design, we tested the use of droplet barcodes to tag all members within a single bacterial aggregate, and found evidence that intrageneric interactions are the most likely to survive strong perturbation. Additional technological development in this field could aid in targeted co-cultivation and add a powerful new tool for interrogating bacterial microenvironments.

3.1 Introduction

The vast majority of microbes on earth exist in biofilms or complex aggregates (67). This spatial structuring can impact community access to resources, protection from antimicrobials, and exchange of metabolic products (21). Physical structuring also reflects ecological interactions between microbial species such as feeding, parasitic, or mutualistic relationships (68). Interest in cell-cell co-localization has developed in parallel with efforts to cultivate a broader diversity of microbial life, since many uncultivated species have reduced genomes and require closely associated partner strains to support their survival (13). Many studies have documented communities in bulk over time or geographic space and inferred relationships between microbes, but high-throughput assays for direct physical interactions remain a technical challenge (69).

Existing techniques to map the micron-scale physical structure of bacteria largely employ variations of flow sorting or microscopy. A common technique to separate and sequence individual cell aggregates involves fluorescence-activated cell sorting followed by target gene amplification and sequencing (30,70). This technique provides sequence-based identification of micro-communities, but is limited by the physical number of wells available for sorting in addition to reagent costs. Fluorescence microscopy is an attractive, low perturbation alternative that has revolutionized the study of biofilm spatial structure, particularly in the oral microbiome (26,71). However microscopic fluorescent probes are still challenging to develop and parallelize, and thus cannot produce high-throughput sequence-based identification of microbial species (36).

We hypothesized that the preservation of microbial spatial structures by hydrogel polymerization, combined with fusion PCR to link DNA fragments between microbial species, could provide a high-throughput, low cost alternative for microscale spatial sequencing. This approach would heavily draw on core molecular techniques from epicPCR, our recent method that physically links genes into individual amplicons within single bacterial genomes (72). Hydrogel bead capture provides a versatile means of fixing spatial relationships within microscopic bacterial aggregates, and can scale down to function in nanoliter emulsion droplets with a benchtop protocol (73). Hydrogel beads encapsulating bacterial genomes can then serve as templates for fusion PCR with novel primer designs. In order to record spatial information within droplet-localized aggregates, primers could be adapted to target rare clades and create cross-species sequence products that all record spatial proximity in the original droplet.

The ability to map spatial structure with next-generation sequence constructs would revolutionize the study of multi-species biofilms, so we tested a variety of emulsion designs to assay cell-cell co-localization in human dental plaque and tongue dorsum biofilms. The easily accessible biofilms that form in the human oral microbiome provide a well-characterized sample source to benchmark experimental perturbation levels and test clade-specific targeting (74). We first tested a clade-targeted design by forming droplet-localized fusions to members of the *Streptocccus* genus and *TM7* phylum. Next, we used a universal droplet barcoding design to probe all bacterial spatial relationships in a suspended sample, and paired this design with synthetic spike-in aggregates in samples from multiple individuals. Our parallelized, sequence-specific findings demonstrate the potential and versatility of this approach.

3.2 Results

3.2.1 Benchtop emulsion designs to capture physical interactions in multi-species biofilms

epicPCR provides a workflow that can be altered to capture microscopic cell aggregates into microscopic polyacrylamide droplets (72). An aqueous sample is mixed with acrylamide and then emulsified into a mineral oil emulsion by aspiration, which reduces shear force relative to vortexing in order to maintain microbial spatial associations. After polymerization, the oil is removed, resulting in polyacrylamide beads, some of which have cell aggregates captured inside. Since the standard workflow generates approximately 500 million polyacrylamide beads, epicPCR permits detecting a very large number of physical interactions between microbes.

The captured cells are subjected to a PCR that fuses together a segment of the prokaryotic 16S ribosomal RNA gene with either a fragment of a clade-specific gene, or droplet-specific barcodes (Fig. 3-1). In a clade-targeted design, the assay targets one clade-specific gene and identifies instances of other bacterial species co-occurring with that clade via a fusion PCR directly analogous to that used in epicPCR (72). This approach is especially useful for targeting rare members of a population that may not produce rich information in an untargeted spatial study. In our droplet-barcoded design, an untargeted all-against-all assay is performed to detect any spatial co-occurrence between any bacteria in the same droplet. This is achieved by amplifying a synthetic 20-mer barcode in place of a targeted gene, and these barcodes are diluted down to a level of less than one barcode expected per droplet. Grouping barcodes in the final sequence data produces groups of sequences that may link between different 16S rRNA gene segments.

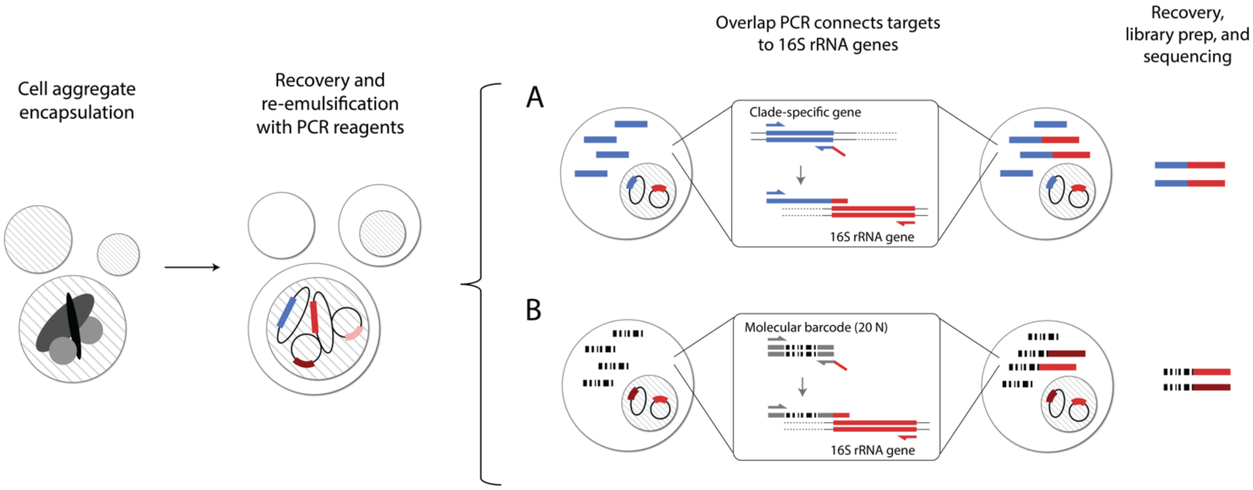
Cell aggregate encapsulation
Recovery and re-emulsification with PCR reagents
Overlap PCR connects targets to 16S rRNA genes
Recovery, library prep, and sequencing
A
Clade-specific gene
16S rRNA gene
B
Molecular barcode (20 N)
16S rRNA gene

Figure 3-1. A schematic depicting our two primary designs for recording microaggregate spatial associations. Biofilms or particles are gently suspended in emulsion droplets via aspiration, then encapsulated in polymerized hydrogels and recovered for bulk lysis. Hydrogels containing aggregate exposed genomes are resuspended with PCR reagents for one of two different primer designs. A) In a clade-targeted design, we amplified a clade-specific gene within droplets, so that only droplets containing that clade would produce double-stranded products with overhangs to the 16S rRNA gene. These overhangs, upon reverse-complementing a bridge primer, would then act as primers on the 16S rRNA gene to form fusions between some member of a clade and any 16S rRNA gene available in the same droplet. B) In a second design, 20mer barcodes located between two constant sequences are added at low dilution and amplified within each droplet. A bridge primer links the

amplified barcode to any 16S rRNA gene available in the droplet, and fusion products are recovered, sequenced, and grouped by barcode to identify spatial relationships.

The molecular details of both the clade-targeted and untargeted barcode approaches are similar. The PCR reaction takes place in emulsion compartments, providing each bead with a discreet reaction compartment that does not exchange molecules with any other reaction. Thus, each fusion will take place between individual genomes that were captured together in a polyacrylamide bead. The fusion of two amplicons in one reaction is achieved using a limited cycle-PCR with primers that initially amplify clade-targeted region or droplet barcode exponentially and the 16S rRNA gene linearly. The limiting 16S rRNA gene primer has an overhang complementary to the targeted segment that will eventually extend to the 16S gene, creating the fusion amplicon. This amplicon is subsequently amplified with an abundance of paired 16S and targeted primers.

To prepare the fusion amplicons into Illumina MiSeq libraries, the product is re-amplified in a nested PCR while including blocking primers that prevent the extension of incompletely amplified products. This PCR also includes Illumina-compatible overhangs that permit subsequent indexing of the library. The fusion amplicons are sequenced on an Illumina platform and any observed interactions between the 16S rRNA gene Operational Taxonomic Units (OTUs) can be processed into an interaction network.

3.2.2 Clade-specific targeting identifies curated spatial partners

We first pursued a clade-targeted design to identify spatial partners of specific taxonomic targets, with the eventual goal of informing co-cultivation attempts for candidate phyla. Our hypothesis was that common, reproducible spatial partners, especially those that occur between multiple species in a clade, may have an auxotrophic relationship that we could exploit in co-cultivation trials. This approach required designing clade-specific primers which landed on shared segments within a multi-gene alignment, in order to amplify a specific clade as well as capture sequence variation within the clade. To benchmark a targeted design for our spatial structure assay, we first targeted the *Streptococcus* genus using primers described in Table B-1. This genus was chosen since it is incredibly well studied and documented in microscopic as well as co-aggregation assays (36,75) (Table 3-1).

We sampled one healthy subject and prepared hydrogel beads using biofilms from both the tongue dorsum and supragingival plaque. Each collection was suspended, and then a negative control strain, *Shewanella oneidensis*, was spiked-in and gently mixed prior to bead polymerization. For each sample site we completed three replicate epicPCR assays linking a targeted region of the *Streptococcus* 23S rRNA gene with any universal 16S rRNA gene co-occurring in the same droplet. Previously validated genus-specific *Streptococcus* primers were derived from (76). The targeted 23S rRNA gene segment was approximately 126bp after the nested reaction, so the 23S rRNA gene sequence was split from the forward reads and analyzed separately from the universal 16S rRNA gene segment.

The qualitative *Streptococcus* pairings that we recovered show strong correspondence with existing co-aggregation literature. The fusion products recovered showed an expected design, with a small number of *Streptococcus* OTUs linking to specific OTUs in the broader phylogeny of the original sample (Fig. 3-2). We observed no fusions to the spiked-in negative control, which appeared at 0.1% relative abundance in both the tongue dorsum and supragigival plaque collections. The recovered putative spatial partners include known co-aggregators with *Streptococcus*, such as *Prevotellaceaea*, *Neisseria*, *Haemophilus*, *Streptococcus*, *Veillonella*, and *Capnocytophaga*. Co-aggregation literature lending support to these recovered interactions at the genus level is listed in Table 3-1.

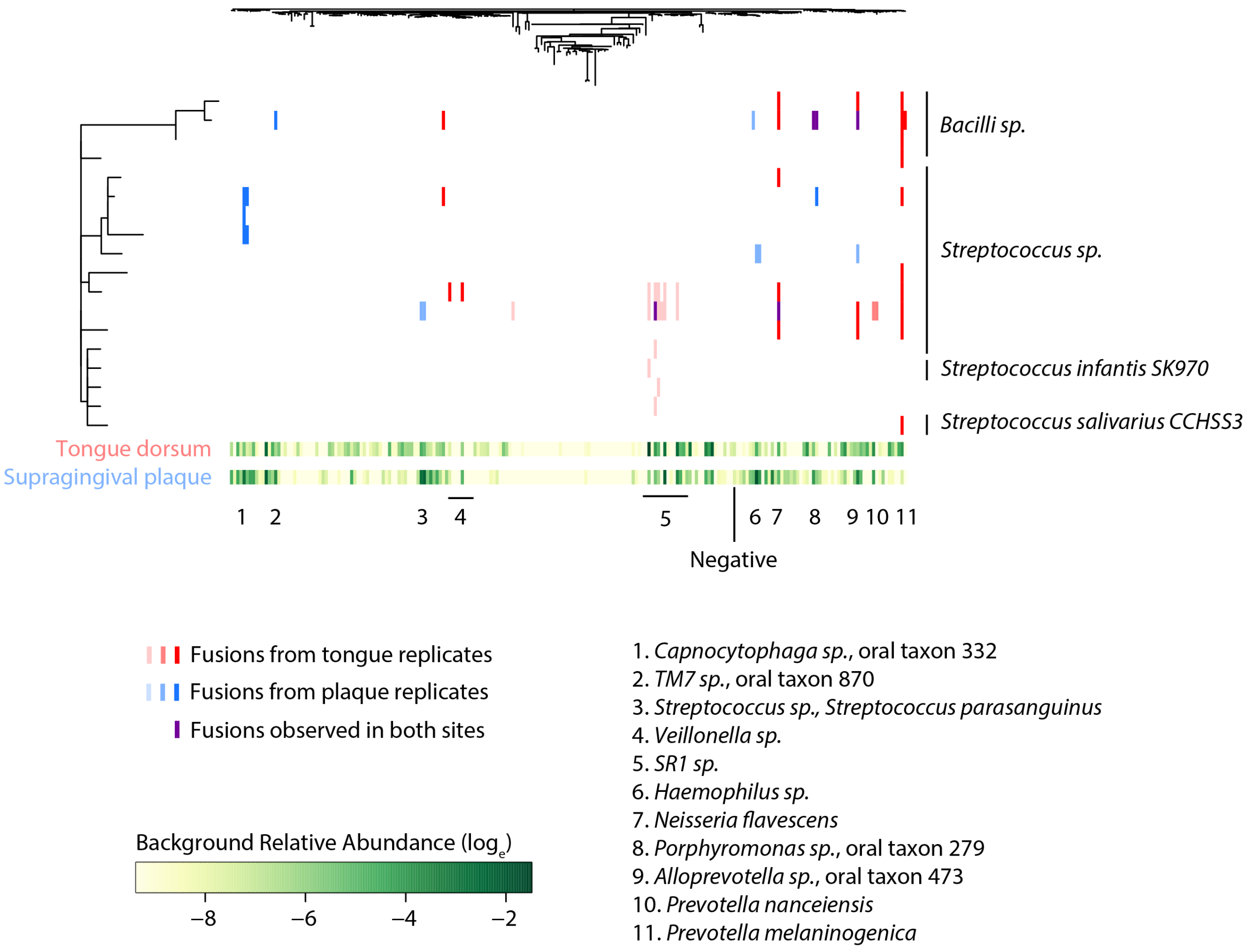

Figure 3-2. Fusion sequences recovered connecting a *Streptococcus* genus-specific 23S rRNA gene segment with universal 16S rRNA. The x-axis tree represents all V1/V3 variable region OTUs recovered across both fusion and bulk DNA sequencing. The y-axis tree represents *Streptococcus* 23S rRNA gene segments recovered. Internal bars show the presence of fusion products connecting between *Streptococcus* and a universal 16S rRNA segment. Background relative abundance for all OTUs is depicted in green below the trees and derives from either the tongue dorsum or supragingival plaque. The numbers below the plot describe the taxonomy of species recovered in fusions with *Streptococcus* 23S rRNA gene segments.

Table 3-1. Co-aggregation literature supporting different recovered genera connected via fusion constructs to the *Streptococcus* genus.

Fusion construct*	Recovered genus	*Streptococcus* co-aggregation sources	Citations
1	*Capnocytophaga*	Cassels and London, 1989	(77)
3	*Streptococcus*	Ruhl *et al.*, 2014	(75)
4	*Veillonella*	Chalmers *et al.*, 2008	(78)
6	*Haemophilus*	Palmer *et al.*, 2017	(79)
7	*Neisseria*	Ruhl *et al.*, 2014	(75)
8	*Porphyromonas*	Maeda *et al.*, 2012	(80)
9	*Alloprovotella*	Schulze-Schweifing *et al.*, 2014**	(81)
10, 11	*Prevotella*	Kolenbrander *et al.*, 1985	(82)

* Numbers correspond to genera identified in Fig. 3-2
** Relative abundance association in dental caries

We also recovered a cluster of epicPCR linkages between *Streptococcus* and the candidate phylum *SR1*, and since *SR1* is currently uncultivable we attempted to establish an *SR1* culture via co-cultivation with different *Streptococcus* species. We designed a bait-based cultivation trial, using a panel of established *Streptococcus* isolates as bait for incubation with *SR1*-enriched samples. A panel of five different *Streptococcus* strains along with a gram – control strain was cultivated and combined with one subject's sputum sample known to carry high *SR1* relative abundance. Despite both microaerophilic and anaerobic passage over a 7 day growth period, we were unable to recover *SR1* in these cultures as determined by a PCR assay (see Section B.1.1 for details).

In a second clade-targeted study we assayed for spatial partners of *TM7*, a rare, poorly studied, and barely cultivated phylum that shows associations with some oral pathologies (13). We generated three replicates from the supragingival plaque of the same healthy subject, confirmed to carry low but consistent levels of TM7 at the sample site. The recovered fusions primarily linked between *TM7* species and other members of the *TM7* phylum, with multiple examples of pairings recovered from different replicates (Fig 3-3). Similar to the Streptococcus-targeted study, the negative control strain was not recovered in the targeted libraries, even when it appeared at >1% relative abundance. Aside from a grouping of inter-phylum fusions, the other clade that was recovered in putative association with members of the *TM7* phylum were species of *Veillonella*, particularly *Veillonella parvula*. *Veillonella* species have been recovered in subcommunities that also include *TM7*, but an exclusive co-culture has not been reported to our knowledge (83).

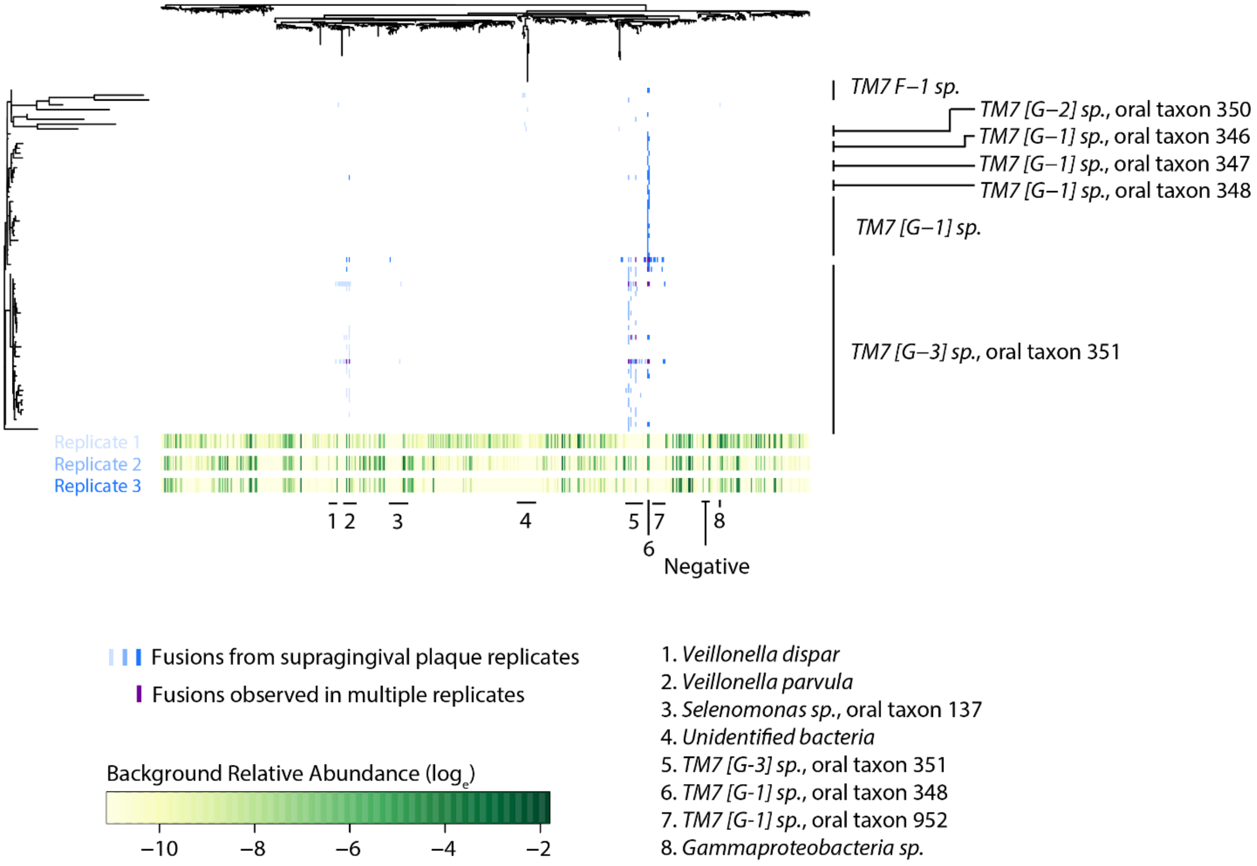

Figure 3-3. Fusion sequences recovered that connect between a targeted *TM7* 16S rRNA segment and the universal V1/V3 variable 16S rRNA gene. The top tree depicts all OTUs recovered from all epicPCR and bulk 16S rRNA gene sequencing. The left tree shows the phylogenetic relationship of targeted *TM7* 16S rRNA gene segments. Bars between the trees show fusion products recovered from three replicate experiments on supragingival plaque, with replicates in different shades of blue and fusions common between replicates in purple. Numbers below the plot are described in the lower right as OTUs recovered in fusions with the *TM7* phylum, and the negative control spike-in is also indicated below the plot.

3.2.3 Molecular barcodes identify enriched spatial connections that dissipate with phylogenetic distance and shear force

In addition to our clade-targeted designs, we implemented an epicPCR design which relied on amplified droplet barcodes linking to any universal 16S rRNA gene segment within a shared droplet. For our first trial, we collected and pooled three supragingival plaque scrapings from one healthy subject and divided it into three replicates that were emulsified at low, medium, and high levels of shear force (characterized by a shift from aspiration to increasing amounts of vortexing). Once hydrogel beads were formed from these different perturbation levels, they were processed identically through an epicPCR protocol that included a 10 pM droplet barcode calculated to load less than one barcode per droplet on average. A similar blank negative control was also processed and sequenced on the same lane, but showed no amplification or significant sequence counts.

When we grouped the resulting sequence data by droplet barcode and then extracted the 16S rRNA gene taxonomies, we found that sequences connected by a common barcode largely fell within closely-related taxonomic groups (Fig. 3-4). We filtered all observed connections by calculating the expected value of observations based on the product of the relative abundances of each OTU pair and the total number of singleton droplet barcodes, defined as a single observation of a unique barcode. Connections that were supported by significantly more barcodes than expected under a Poisson model are depicted at a significance cutoff of 0.01 and 0.001 (Fig. 3-4A and B, respectively). The remaining observations primarily link sequences within closely-related taxonomic groups, with the only replicated cross-phylum interaction showing a connection between *Leptotrichia* and *Veillonella* species. Within clades, increasing shear force reduced the number of total connections observed and also reduced the phylogenetic distance between observations. At the highest level of shear force, only a few strongly connected OTU pairs retained significance within the *Veillonella* and *Neisseria* genera.

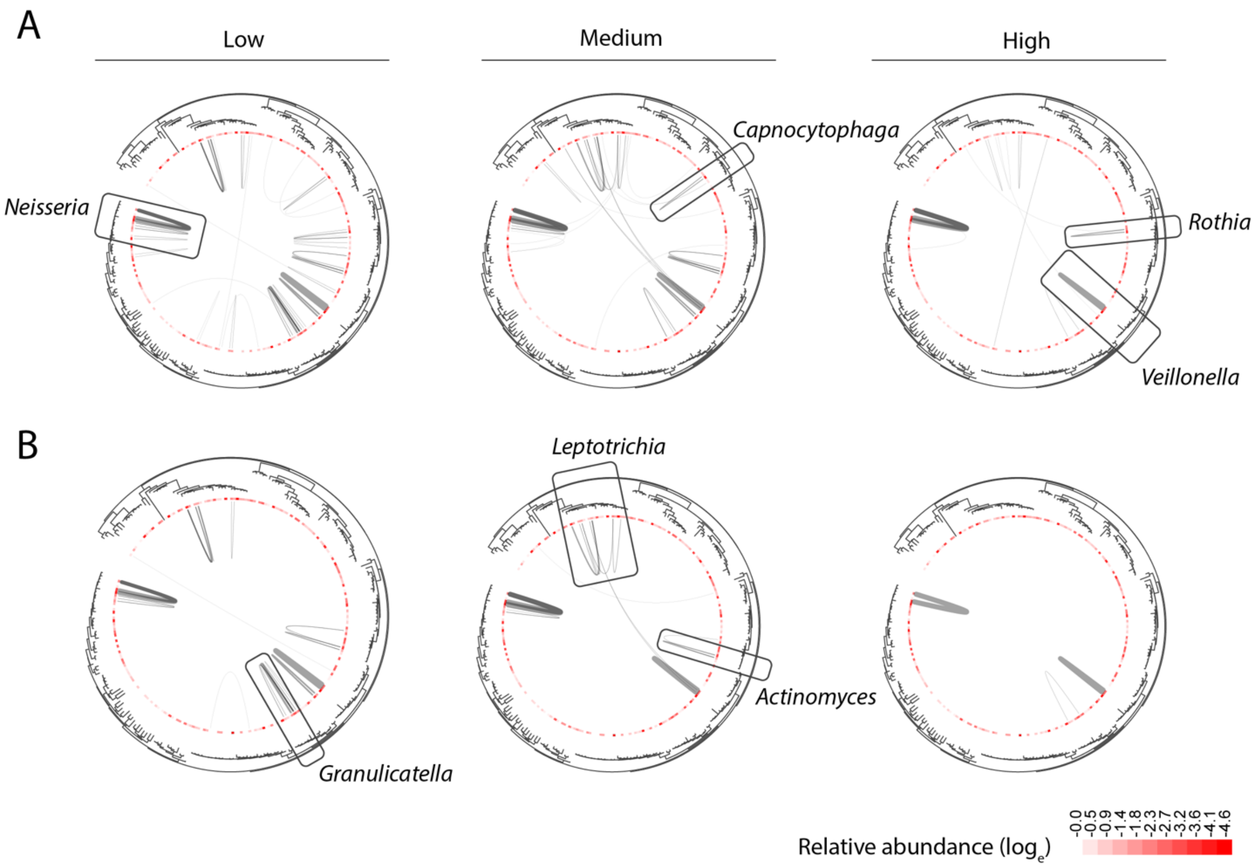

Figure 3-4. Initial molecular barcoding trial shows conserved associations despite perturbation. A) Connections recovered with a Poisson cutoff of $p < 0.01$. B) Connections recovered with a Poisson cutoff of $p < 0.001$. In both panels, genera with significant connections are indicated with boxes, and these labels apply to all trees in the figure. Background relative abundance of the OTUs is shown in red on a log scale.

3.2.4 Barcoding replicates in multiple individuals support stronger physical aggregation between closely related taxa

Due to our reproducible results from oral barcoding in a single individual, we scaled the study to sample multiple individuals and add additional controls. Our experimental design is summarized in Figure 3-5A, showing sample collection from four individuals and hydrogel encapsulation at three levels of shear force with duplicates for each. We also designed a positive control spike-in to accompany our negative control culture of *S. oneidensis*, both of which were spiked in at low abundance in every replicate. Our positive control consisted of synthetic aggregates between *E. coli* and *B. subtilis* produced by treating 1:1 combinations of dense cultures with glutaraldehyde (Fig. 3-5B). Replicates with no cells as well as only spike-in cells were generated in parallel, and all samples were barcoded using 10 pM droplet barcodes targeting the 16S rRNA gene V4 variable region as described in Figure 3-1B.

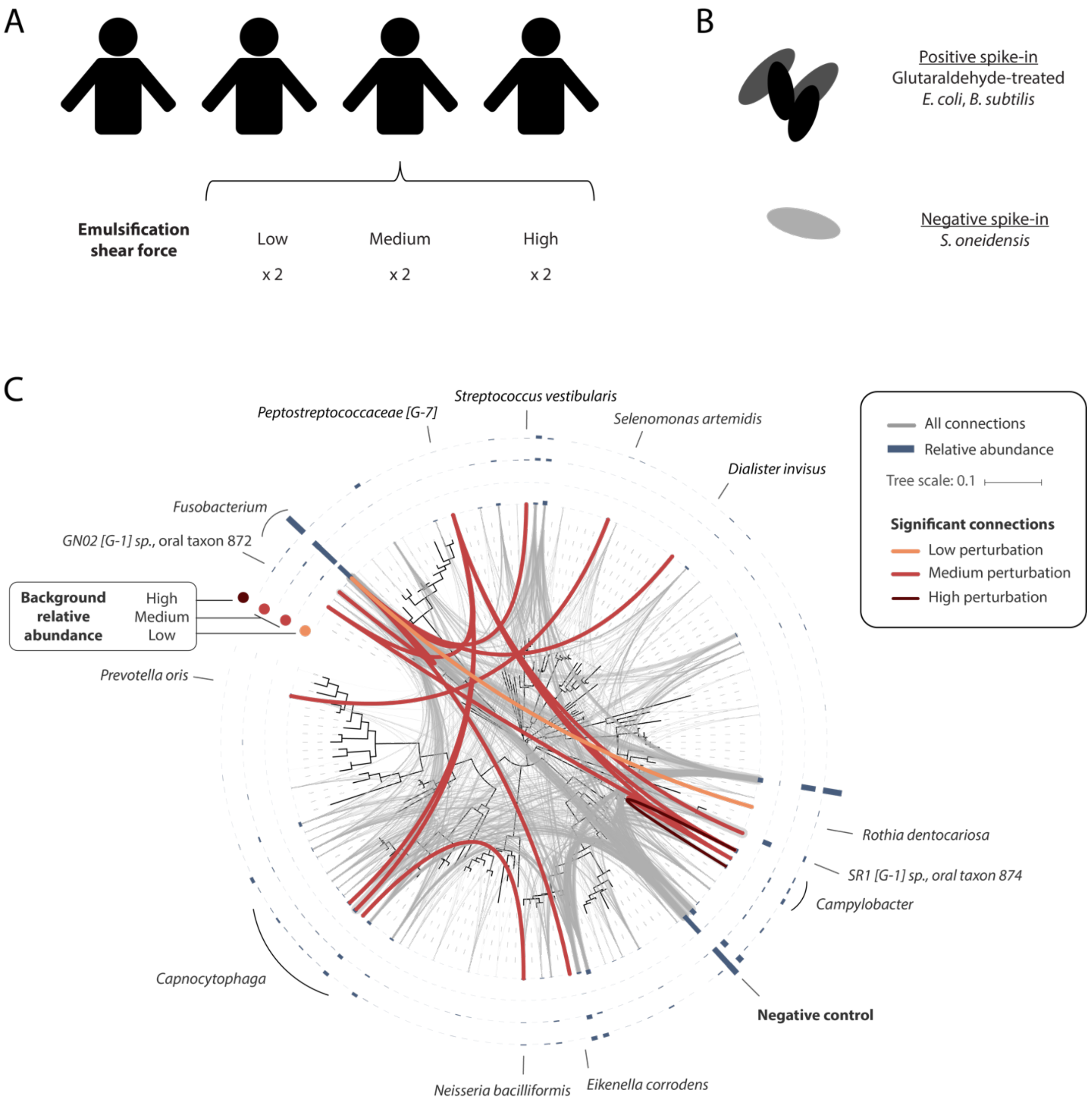

Figure 3-5. Replicate droplet barcoding in supragingival plaque shows the greatest recovery of interactions at medium levels of perturbation. A) Four subjects provided supragingival plaque, which was emulsified at three different levels of shear force with a duplicate at each level. B) Immediately prior to hydrogel encapsulation, we spiked in positive and negative control strains into each sample. Our positive control consisted of glutaraldehyde-treated *E. coli* and *B. subtilis* cells, chosen at a density that provided the most aggregates per μl. Our negative control was a fresh culture of *S. oneidensis*. C) Replicate droplet barcoding in supragingival plaque from subject E. Data presented shows four samples with no significant connections to the negative control strain. Blue bars represent relative abundance of OTUs in the four replicates, with increasing perturbation in the outermost graphs. Grey lines show the raw recovered pairings, weighted by the number of droplet barcodes supporting the pair. Red lines show positive interactions that were significant over a Poisson model ($p < 1e-3$) for different replicates.

We clustered the resulting amplicon data by droplet barcode and identified cross-species interactions supported by more droplet barcodes than expected under a perfectly mixed Poisson model. We first identified all multiplet barcodes that fused to more than one OTU (Fig. B-1), then determined if the number of barcodes supporting a unique pairing was higher than expected by chance considering the two individual relative abundances. The number of significant interactions recovered, either positive or negative, are listed in Table B-2. Despite the weak amplification of glutaraldehyde-fixed positive controls, the *S. oneidensis* cultivated negative control still provided a gating factor for samples with no evident crosstalk after Poisson filtering (Fig. B-2, Table B-2).

Samples from one subject produced four replicates, including one at each perturbation level, with no sign of negative control crosstalk after Poisson filtering. The positive interactions from this subject are presented in Figure 3-5, where the majority of recovered interactions were from one replicate with intermediate perturbation. The largest number of connections to a single genus involved the genus *Fusobacteria*, a well-known late-stage colonizer with broad profiles of intergeneric co-aggregations (74). We also recovered within-genus connections in the *Fusobacteria*, which is supported by co-aggregation studies (84). At the highest perturbation level, we only recovered an intrageneric interaction within the *Campylobacter*, similar to our intrageneric results at high perturbation in Figure 3-4. The remaining intergeneric interactions are less well-characterized, and may indicate possible co-aggregations with candidate phyla including *GN02* and *SR1*.

3.3 Discussion

The ability to assay microbial community structure at a scale relevant to individual cells could shed light on colonization, auxotrophic relationships, and survival in complex biofilms. We attempted alterations of our previously described method, epicPCR, in order to capture and link DNA from micron-scale aggregates of dispersed human dental plaque. Our clade-targeted approach amplified highly specific spatial partners for members of the *Streptococcus* and *TM7* clades. Titrating in droplet barcodes as an indication of spatial relationships generated rich data from multiple individuals, and showed that increased sample handling reduced the recovery of intergeneric associations. Despite the promise of the approach, careful controls for cell loading and stochastic association were necessary and require further development.

Key technical advances from this work included a comparison of targeted vs. untargeted primer designs for spatial sequencing, and an exploration of shear force conditions to disrupt and suspend biofilms. Our three-primer, clade-targeted designs showed the highest level of reproducibility, overlap with literature, and sensitivity to rare taxa, but new implementations should be carefully designed to capture species-level diversity within the clade-targeted site. Our untargeted approach generated richer sequencing libraries and more information per sequencing run, but showed a low signal to noise ratio and more sensitivity to initial loading conditions. For shear force conditions, we found that typical emulsification conditions (> 1 min vortex, stir bar treatment) proved too disruptive for biofilm separation. As an alternative, we present variations of aspiration with small pulses of vortexing, and find that a comprehensive aspiration plus a pulse of stronger shear force generated the greatest number of molecular interactions.

Overall our approach to biofilm sequencing via emulsified aggregates introduced some challenges for sampling handling and control design. An initial droplet encapsulation to preserve spatial structure may be better suited for naturally suspended aggregates rather than bulk biofilms, such as freely suspended and colonized particles in aquatic communities. If biofilms are the target community, it could be useful to perform an initial bulk hydrogel capture or fixation to preserve the initial structure, followed by shearing and droplet suspension. We also found that adding controls was non-trivial, and recommend new approaches particularly for positive control design. While glutaraldehyde-fixation visually aggregated cultured cells, the fixation and storage process dramatically reduced downstream amplification. Exploiting naturally adhesive strains such as *Caulobacter crescentus* could be a useful alternative (85,86).

The general approach presented in this chapter and particularly the molecular biology shows promise, but may benefit by porting the method into a bulk hydrogel design. Bulk hydrogel chemistry is a powerful tool to translate micron-scale spatial information into molecular readouts, and naturally involves less initial sample perturbation. One prominent example of the potential of the field is Fluorescence In Situ SEQuencing (FISSEQ), which uses a modified polyethylene glycol matrix to prevent cDNA diffusion and allow *in situ* sequencing of tissue cross-sections (87). Bulk hydrogels have also been used for single-cell microbial sequencing, and the adaptation of this technique to microbial spatial sequencing would be a natural extension (20). Future efforts should also focus on translation of high-throughput co-localization data into improvements in high-throughput co-cultivation or live cell assays, in order to translate this novel data type into functional insights.

3.4 Materials and Methods

3.4.1 Sample collection

Human supragingival plaque or tongue dorsum samples were collected with sterile instruments and suspended in 130 µl PBS buffer, then stored on ice. When described, approximately 1 million *Shewanella oneidensis* cells (strain MR-1, ATCC 700550, Manassas, VA, USA) were spiked in from liquid culture. Glutaraldehyde-prepared positive control cells (a mixture of *Escherichia coli* and *Baccilus subtilis*) were added to the final barcoding experiment as described in Section B.1.2. Aliquots of 30 µl of each suspended cell sample were immediately ported into the epicPCR workflow. Any remaining material not used in the epicPCR workflow was combined 1:1 with 50%

glycerol and frozen at -80C for downstream DNA extraction and background library preparation (details in Section B.1.3).

3.4.2 Cell aggregate encapsulation for epicPCR

Each cell suspension was mixed into a solution with 9.4% acrylamide, 0.25% N-N'-bisacryloylcystamine, and 0.98% ammonium persulfate for a total volume of 255 µl. This solution was gently swirled to mix and then combined with 600 µl of mineral oil containing 4.5% Span 80, 0.4% Tween 80 and 0.05% Triton X-100 (52). The samples were emulsified according to various protocols, producing approximately 500 million individual aqueous droplets in oil. We then added TEMED to an aqueous concentration of 8.9% and emulsified again to distribute the polymerization catalyst. The three protocols used for emulsification included a light perturbation (10x aspiration, TEMED addition, 10x aspiration), medium perturbation (10x aspiration, 2 sec vortex, TEMED addition, 10x aspiration, 2 sec vortex) and strong perturbation (10x aspiration, 10 sec vortex, TEMED addition, 10x aspiration, 10 sec vortex).

After 1.5 hours of polymerization at room temperature, polyacrylamide beads were recovered by centrifuging 1 minute at 13 000 *g*. Excess oil in the upper phase was discarded and replaced with 800 µl diethyl ether. The sample was mixed, then the ether was removed from the upper phase and replaced with sterile, nuclease-free water. After mixing, another centrifugation identical to the first again distributed the polyacrylamide beads to the lower phase and allowed the removal of residual oil in the upper phase. Residual oil removal, replacement with nuclease-free water, and centrifugation were repeated until the upper phase was clear. Beads were treated with lysis reagents and filtered through 100 µM size-selection mesh filters as described in Section B.1.4. Beads were stored at 4°C in darkness until resuspension for epicPCR reactions 1-4 days later.

3.4.3 epicPCR library preparation

epicPCR was performed using primers to link together bacterial 16S ribosomal RNA gene sequences with droplet-specific barcodes. Each PCR mixture included 46.5 µl of lysed and filtered polyacrylamide bead suspension, a high concentration of Phusion Hot Start Flex DNA Polymerase (NEB, Ipswich, MA, USA), and a combination of forward, reverse, and bridging primers to link together targeted genes (primer sequences and concentrations in Table B-1). Specifically, the reactions included 1X Phusion HF buffer, 1 mM $MgCl_2$, 250 μM each dNTP, 50 ng/μl BSA, 0.2% (v/v) Tween 20, and 0.16 U/μl Phusion Hot Start Flex. The 100 μl PCR master mix was emulsified into 900 μl of mineral oil previously combined with 4% ABIL EM 90 (Evonik, Mobile, AL, USA) and 0.05% Triton X-100 (v/v, molecular biology grade, EMD Millipore, Billerica, MA, USA). The emulsion was formed by vortexing 1 minute at full speed, and the suspension of beads into individual reaction compartments was enhanced by adding 3-4 glass beads (2 mm diameter, Andwin Scientific, Schaumburg, IL, USA) into each tube.

After emulsification the emulsion was partitioned into PCR wells in 50 μl aliquots and cycled on a PCR machine (94°C 30 sec, 33 cycles of (94°C 5 sec, 52°C 30 sec, 72°C 45 sec), 72°C 5 min, 10°C hold). The annealing temperature was increased to 60°C and the extension time decreased to 15 sec for both clade-targeted assays. Following amplification, the emulsions were pooled and 1 mM EDTA was added to inhibit any additional polymerase activity. The aqueous phase was recovered with two diethyl ether extractions, one ethyl acetate extraction, and an additional two diethyl ether extractions. The extracted PCR mix was purified using AMPure XP beads (see A.1.4 for details).

To prepare the fused amplicons into a sequencing library for the Illumina MiSeq platform, we performed a nested PCR that included blocking primers to suppress unfused partial constructs. We used nested primers described in Table B-1 which carried Illumina overhang sequences at the 5' ends, and these were loaded at 0.3 μM each. Blocking primers which anneal to unfused bridge overhangs were added at 3.2 μM each (45,88). A nested qPCR, to determine a minimal cycling threshold, and replicate nested PCRs, to reduce jackpot effects, were completed under the conditions described previously (72). Nested PCRs were performed in quadruplicate for each sample and included 1X Phusion HF buffer, 200 μM each dNTP, 0.02 U/μl Phusion Hot Start Flex DNA Polymerase, and the primer concentrations described in Table B-1. Amplicon libraries were purified with AMPure XP beads and indexed into final Illumina libraries by performing an 8 cycle amplification using indexing primers (see B.1.5). All libraries were sequenced on Illumina MiSeq platform with 250 bp (clade-targeted design) or 300 bp (droplet barcoded design) paired-end reads.

3.4.4 Sequence data processing

Data pre-processing included paired-end read joining, quality filtering, and primer checking and removal. For barcoded samples we used PEAR v0.9.10 to complete paired-end read joining using default parameters, and continued with only successfully joined constructs (89). For clade-targeted samples we processed the forward and reverse reads separately. Reads were quality-filtered with usearch v9.2 using the –fastq_filter flag and parameters –fastq_minlen=100, –fastq_maxee_rate=0.01 (63). Reads were demultiplexed and converted to fasta file format using custom scripts available in a publicly accessible jupyter notebook (https://github.com/sjspence/plaque_barcoding/blob/master/jupyter/OM8_pipeline.ipynb). Finally, we confirmed the correct primer structure and removed primers from the sequences using custom scripts in the package epicBarcoder (https://github.com/sjspence/epicBarcoder). Additional details on dereplication and denoising are available in Section B.1.6.

We tailored OTU calling and used different databases for each amplified segment under study. For the two clade-targeted studies, we used the HOMD 16S rRNA gene database v14.5.p9 with the *S. oneidensis* sequence aligned and appended (90). This formed the reference for our targeted V1/V3 variable region segment on the reverse reads. We assigned the targeted *Streprococcus* 23S rRNA gene segment taxonomy with the SILVA large-subunit database v123.1 (61). The two barcoding studies also relied on the HOMD 16S rRNA gene database, and included aligned and appended representative sequences for the *E. coli* and *B. subtilis* positive spike-in controls. Background relative abundance information was calculated from background library preparations or singleton barcode abundances.

Each collection of amplified targets was grouped into representative sequences, aligned, and built into a phylogenetic tree for visualization. In order to reduce noise from the clade-targeted libraries, which required higher numbers of PCR cycles, we employed a 97% sequence identity clustering with usearch v8 (–cluster_fast –sort length –centroids –id 0.97). In order to assign universal barcode sequences more closely to their representative taxa, we grouped those sequences by common HOMD assignments with custom scripts. All reads were aligned with SINA v1.2.11 under default parameters, and including the –ptdb flag to either the SILVA v128 small- or large-subunit reference alignment databases (SSURef_NR99_128_SILVA_07_09_16_opt.arb, LSURef_128_SILVA_20_09_16_opt.arb) (62). Trees were constructed using FastTree v2.1.7 with the –nt and –gtr flags (47).

Chapter 4 Whole genome sequencing of deep branching strains shows evolution and exchange across a contaminated watershed

Sarah J Spencer, Alex B Aaring, Austin Hendricks, Jon Penterman, Romy Chakraborty, Eric J Alm

The work in the chapter presented below constitutes a manuscript in preparation for submission.

Abstract

Microbial communities in groundwater ecosystems are inherently difficult to study due to sampling challenges paired with highly dynamic microenvironments. A number of recent efforts have characterized and mined the vast bacterial diversity within groundwater via amplicon surveys and shotgun metagenomics, but these data sources cannot reveal dynamic or historical changes without high-resolution sampling. Here we aimed to identify recent microbial adaptation within groundwater sites by capturing evolutionary signatures from isolate whole genome sequencing. We recovered 139 *Pseudomonas* isolates which grouped into 15 strains, and found that each strain contained isolates from multiple sampling sites across a broad geographic region. SNP analysis confirmed that recent mutational changes were impacting strains in individual sampling sites, despite a fast-flowing aquifer, and some SNPs in a two-component system showed evidence of positive selection. We also searched for gene gain and loss within strain isolates, and found evidence of transcriptional gene excision within otherwise clonal isolates at a single location. Since isolate sequencing within environmental strains can reveal recent adaptive changes, it is a powerful tool to understand the impact of groundwater perturbations through the lens of microbial genomes.

4.1 Introduction

Groundwater microbiology faces challenges from massive microbial diversity, difficulty of sampling, and highly dynamic environments. Microbial diversity in groundwater is high relative to other environments and captures a multitude of unknown or candidate phyla, some of which dominate these communities (91,92). With so much unknown phylogenetic diversity comes a variety of novel metabolisms, many of which developed in response to human impact, such as chromium reduction or phenoxy herbicide degradation (93,94). This phylum-level and metabolic complexity arises from highly complex and dynamic geochemistry that changes along vertical and horizontal gradients in the environment (95,96). These transects are also highly dynamic at each sampling point, and show rapid microbial shifts based on rainfall events or even diurnal cycles which can only be captured by continuously flowing sampling lines (97). Despite these challenges, groundwater communities are critical to investigate as they host a huge fraction of prokaryotic life on earth and provide a first line of defense against human industrial contamination (98).

In order to tease apart the complex metabolisms present in different members of groundwater ecosystems, as well as discover new entirely new phyla, many efforts have pushed to construct individual bacterial genomes from metagenomic or isolate data. Thousands of aquifer genomes have been constructed from bulk metagenomic data, and collectively they encompass the diversity and common features of the candidate phyla radiation, as well as revealing metabolic handoffs in shared sampling sites (92,99). Another approach has been to construct population genomes from shotgun-sequenced enrichment cultures of groundwater, in order to enable more confident metabolic reconstruction (100). These efforts join a steady stream of individually published draft or complete bacterial genomes from groundwater isolates, which are often sequenced due to harboring unique metabolisms or extensive use in model systems (93,94,101).

While groundwater genome reconstruction efforts have effectively described tremendous genomic diversity, this diversity captures millions of years of evolution rather than changes occurring on a timescale relevant to human environmental impact. The latter requires studying genomes from closely related species or strains to detect recent adaptation and spread (102). Whole genome sequences from the same species or even genus are rare in groundwater ecosystems. One serial enrichment and isolation effort from a deep aquifer recovered three genomes that fell into different taxonomic classes (103). Another study used single amplified genomes to recover four members of the *Pedobacter* genus from aquifer sediment, but completed all analysis relative to other bacterial groups (104). Recently, a strain sequencing effort from surface water pools produced whole genome sequences from closely related strains within the *Ensifer* and *Sinorhizobium* genera, but these data were mainly used to support a change in genus classification (105). To our knowledge, no systematic isolation effort from groundwater has recovered multiple isolates from closely related strains for evolutionary analysis.

Here we present a collection of isolate draft genomes which group into distinct strains, and confirm evolutionary signatures within individual sampling sites despite a rapid flow environment across a groundwater contamination gradient. In general, we found that strains are distributed between wells that are kilometers apart, and individual sampling sites maintain consistent strain diversity. We performed SNP variant detection on 15 strains containing over one hundred isolate genomes within the *Pseudomonas* genus, and reveal one strain group with strong adaptive signatures in a two-component signaling system, largely originating from one sampling location. We also identify putative gene loss affecting otherwise clonal members of a single groundwater site. These findings on adaptation and diversity were derived from one genus of an otherwise highly complex community, so further high-throughput isolation and whole genome sequencing efforts in different clades could reveal novel biology and recent adaptive pathways in groundwater ecosystems.

4.2 Results

4.2.1 Isolates recovered from a contaminated aquifer span a broad phylogenetic range and deeply sample the abundant *Pseudomonas* genus

We reanalyzed the 16S rRNA gene amplicon data from a survey spanning 100 aquifer wells across a transect of the Oak Ridge Watershed in TN to identify genera with high species diversity and high relative abundance in multiple wells (96). This sampling site has undergone heavy metal and uranium contamination due to leached material from lowly contaminated water used in past nuclear processing. An updated operational taxonomic unit (OTU) table from the site was generated with ecologically-informed sequence clustering and over 26,000 OTU's were recovered in total, but only 1,170 showed relative abundance > 1% in at least one sample (106). Of these, the *Pseudomonas* genus contained the second most OTU representatives out of all classified genera, after *Nitrospira*, and many of the *Pseudomonas* OTUs appear in multiple samples and wells across the site (Fig. 4-1A, Fig. C-1).

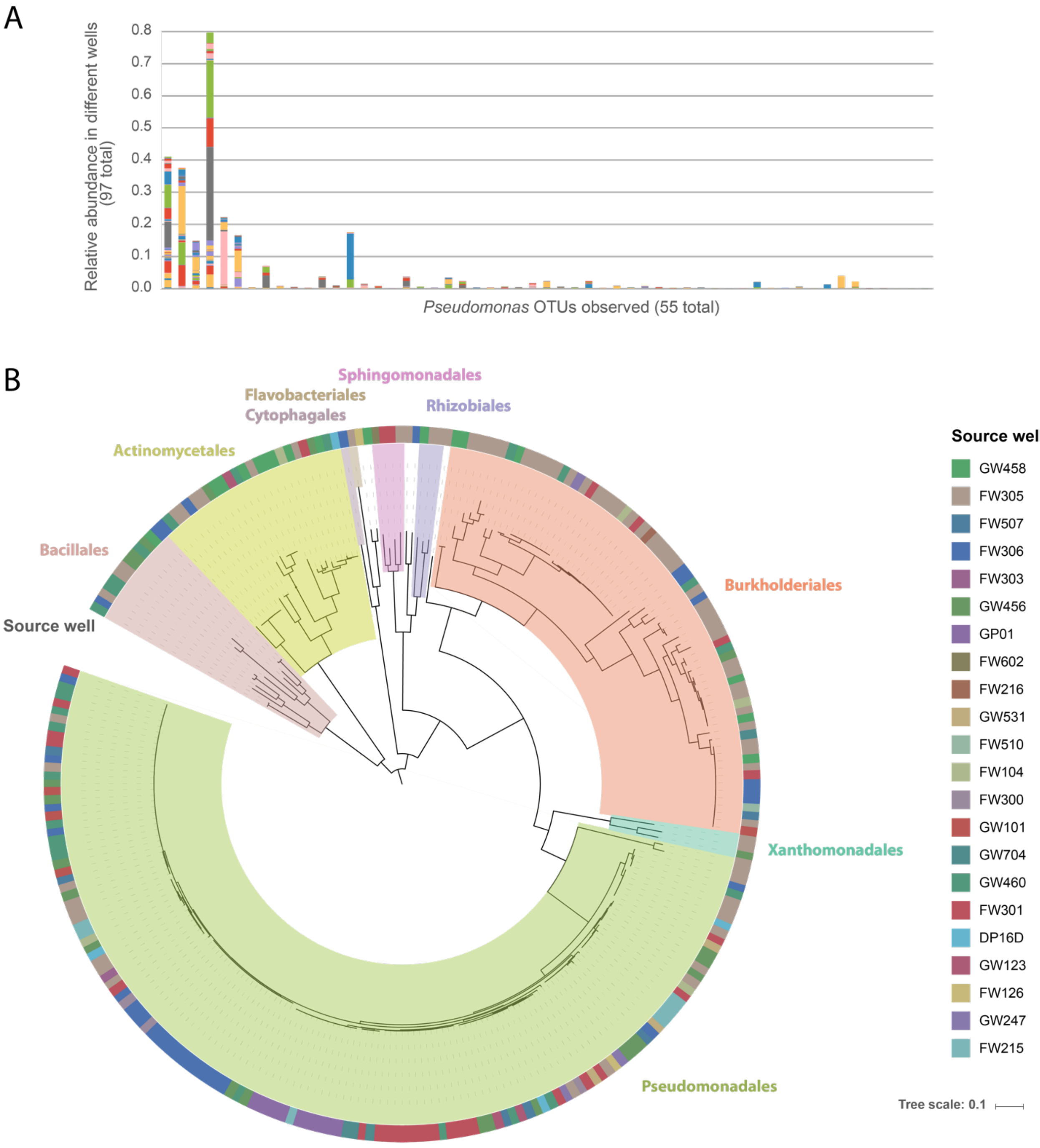

Figure 4-1. Isolation and whole-genome shotgun sequencing targeted toward the *Pseudomonas* genus. A) 16S rRNA gene relative abundance for OTUs within the *Pseudomonas* genus, with colored bars representing relative abundance in different samples from 97 wells surveyed in **(96)**. B) Complete diversity of high quality genomes displayed via a maximum likelihood tree of masked

AMPHORA protein alignments. Phylogenetic order is shown in colored arcs over the tree branches, and the geographic source wells for each strain are displayed in colored bars on the outer circle.

We then cultivated and isolated strains from a subset of the geographic sites in the 16S rRNA gene survey, using an array of sparse and rich media as well as both aerobic and anaerobic conditions for cultivation (96). Source wells spanned a broad geographic area, different contamination levels, and different time points (Fig. C-2). As expected, our isolation recovered a large number of species from the *Pseudomonas* genus due to high relative abundance and prevalence across the site, as well as broad amenability to cultivation. This genus is also relevant to the site due to roles in nitrate reduction and potentially uranium fixation, and boasts extensive metabolic characterization relative to other isolates from the region (101,107). We also recovered an equal grouping of non-*Pseudomonas* isolates spanning the broader diversity in the wells, and specifically generated a sizable collection from the order *Burkholderiales*. Within the *Pseudomonas* genus as well as other closely related clades, isolates were often recovered from a broad array of source wells.

An automated, low-volume Nextera protocol generated economical, high coverage whole-genome data. We processed 288 isolates in a single 384-well plate using low-volume Nextera reactions, resulting in all but one sample with more than 500,000 reads and a strong representation of unique 20mers in the forward reads (Fig. C-3). Libraries showed an average of 29% read duplicates, likely due to the decreased input material, but grouped towards low duplication levels (Fig. C-4). We recovered 265 de novo assembled genomes which showed > 95% completeness and < 10% contamination in a checkM marker gene summary (108). For each of these, we recovered and concatenated AMPHORA protein sequences and constructed a maximum likelihood tree to depict accurate phylogenetic relationships among these genomes along with their source wells (Fig. 4-1B) (109).

4.2.2 *Pseudomonas* species and strains actively exchange between disparate geographic locations

Within the *Pseudomonas* genus, we selected strains for independent downstream analysis in order to identify new, rather than ancient, adaptations to the recently contaminated environment. The *Pseudomonas* genus alone contained 139 high-quality de novo assembled genomes that sampled a large number of deep branches within the genus, segregated by an order of millions of years of evolution (Fig. 4-2A). We used a concatenated AMPHORA2 maximum likelihood tree to select subgroups within the genus separated by fewer than 1/1000 AMPHORA protein substitutions as candidate groups for strain-level analysis. These strain subgroups were also apparent at the nucleotide level, and cleanly separated into hierarchical clusters of nucleotide substitutions per site in the AMPHORA alignments (Fig. C-5). Even with the high ribosomal gene similarity, genomes within each strain varied between 69% and 96% core genome percentage determined by whole genome alignment (Fig. 4-2B). Most of the strains labeled with letters A-O below represent unknown species and could only be confidently identified to the genus level.

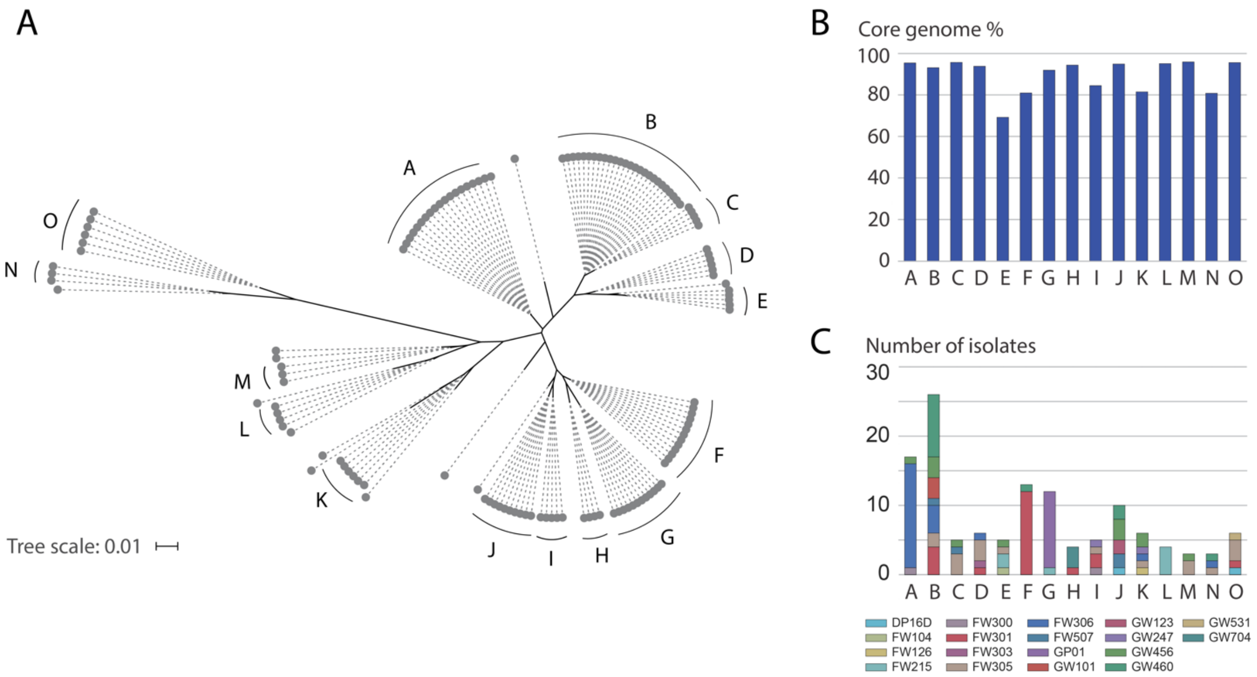

Figure 4-2. *Pseudomonas* isolate genomes separate into deep-branching strains. A) An unrooted tree of *Pseudomonas* AMPHORA2 masked protein alignments. Closely related isolates were grouped into strains indicated with letters, and these strains were used for downstream read mapping and SNP calling. The tree scale is the number of protein substitutions per site in the AMPHORA2 concatenated protein alignment. B) Core genome percentage calculated for each strain based on alignment to a random reference selected from the subgroup. C) The number of isolates recovered within each strain, with colored bars indicating the source well for each isolate.

Similar to the OTU analysis in Figure 4-1A, we enumerated how many wells each strain had cultivable representatives in. Strains were most commonly sourced from four independent sampling sites, although the majority of isolates were typically derived from one of the sites (Fig. 4-2C). This result is similar to the 16S rRNA amplicon data because we do observe physical distribution of strains across the region, similar to the physical distribution of *Pseudomonas* species and counter to the hypothesis that strains would be unique to each well. Another commonality is that strains often have one dominant location, similar to one or two dominant sites observed for each OTU in terms of relative abundance. Finally, in both datasets we recovered extensive overlap of source sites between different species or strains, supporting a model of conserved diversity even at fine-scale taxonomic resolution.

4.2.3 Recent mutation in members of a two-component signal transduction system demonstrates *in situ* adaptation

Within each *Pseudomonas* strain, we then tested if we could detect signatures of evolution or even adaptive mutation in individual sampling sites. We aligned reads from each isolate within a strain to a combined scaffold assembled from all the isolates within the strain, searching for mutations accumulated within the past 50-100 years and thus aligned with human impact on the environment. Based on strict base quality, read alignment quality, and base coverage cutoffs, we recovered high-quality SNPs differentiating isolates from a strain and constructed a SNP phylogeny. Some strains, such as strain G in Figure 4-2A, generated no high-confidence SNPs and appear highly clonal on a nucleotide level. Others, such as strain B, contain 39 high-confidence polymorphisms, although none show obvious positive selection via deviation from expected dN/dS ratios.

Strain B was the only *Pseudomonas* strain to contain multiple non-synonymous mutations in individual protein-coding genes, and these genes grouped into a probable pathway under selection (Fig. 4-3A). Two of the genes with multiple non-synonymous mutations were automatically annotated as the *barA*/*uvrY* two-component signaling system, a pairing of a membrane-bound histidine kinase and a cognate DNA-binding response regulator. When we performed blastp against the NCBI non-redundant protein database, the closest match for the gene annotated as *barA* is a hybrid sensor histidine kinase/response regulator in *Pseudomonas mandelii*. Likewise, the closest match for the downstream gene initially annotated as *uvrY* is a DNA-binding response regulatory in the *NarL*/*FixJ* family, also identified in *Pseudomonas madelii*. Both genes share conserved superfamily and family domains with the *gacS*/*gacA* two-component system, the technical homologs to *uvrY* and *barA* in *Pseudomonas* species. Significantly, the *gacS*/*gacA* two-component system forms a critical upstream regulator of external enzymes, e.g. lipases, and siderophore production in a variety of *Pseudomonas* species (110–112).

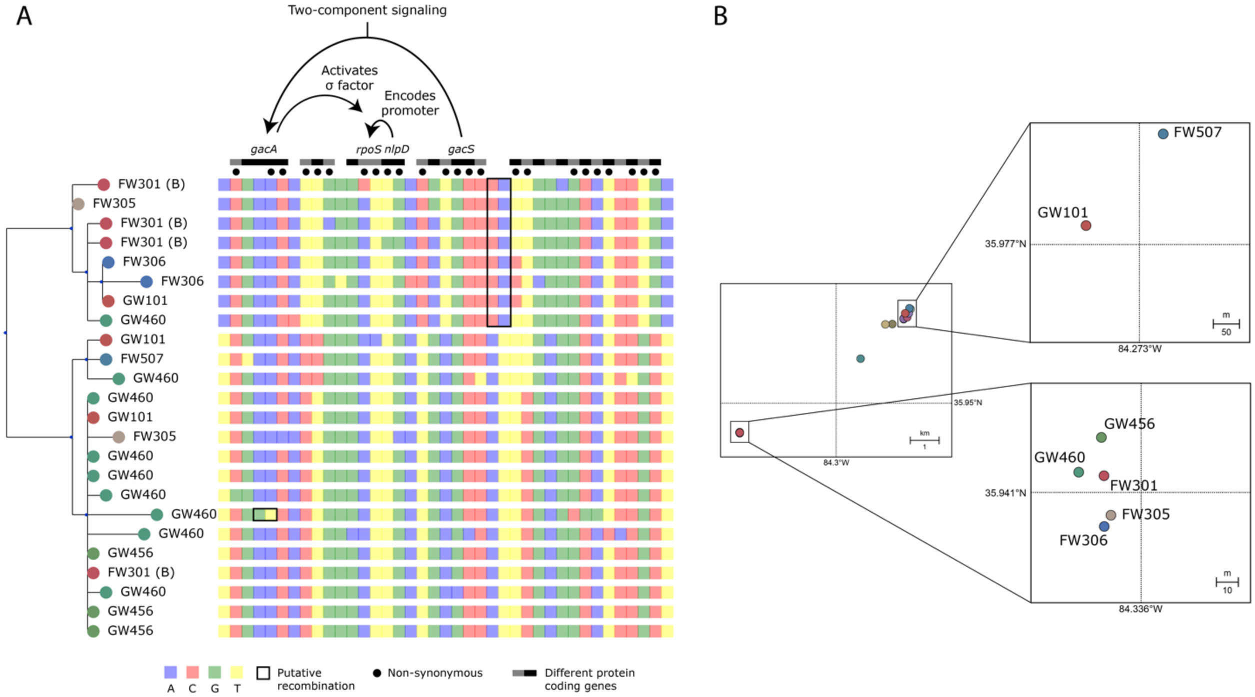

Figure 4-3. SNPs identified in one subgroup of *Pseudomonas* strains disproportionately impact the *GacS*/*GacA* two-component signaling pathway. A) A tree constructed from identified SNPs in strain B has leaves colored by source well, and the letter B appended to those recovered after short-term cultivation in a bioreactor. Each column represents an identified SNP, with different protein-coding genes marked as alternating black and grey bars, and non-synonymous changes marked with a circle above the column. SNPs occurring in the same isolate and locus are marked as putative recombination sites. Above the plot, proteins which share a pathway or regulation are described with arrows. B) The geographical locations of source wells for this subgroup of strains, with zoomed sections in the denser collections of northern and southern wells.

The identification of multiple mutations in the sigma factor gene *rpoS* and the lipoprotein *nlpD* is also relevant because these two genes are transcriptionally connected and likely become activated downstream of the two-component signaling activity. In the homologous *barA*/*uvrY* two-component system, the response regulator *barA* is required for the exponential induction of the *rpoS* sigma factor, which is tied to the bacterial response to hydrogen-peroxide stress (113). The primary promoter for *rpoS* in *E. coli* is contained within the *nlpD* gene, an outer membrane lipoprotein, and the synteny of these genes is maintained in the strain B *Pseudomonas* isolates indicating broad conservation (114). The two strain B *nlpD* SNPs fall within 920 bp of the *rpoS* ATG codon, a region confirmed to have multiple sites with promoter activity in *Pseudomonas putida* (a rhizosphere isolate), although the SNPs likely do not impact the primary promoter which shows cross-species conservation approximately 400 bp upstream of the *rpoS* start codon (114–116).

In addition to studying the gene content and SNP locations, we wanted to understand the accumulation of these SNPs relative to each other and to the sampling region. We identified possible sites of recombination apparent in one GW460 isolate *gacA* gene, with two high-confidence variants carried together on the same locus. We also identified an intergenic region with two co-localized variants carried by a group of isolates from different sites, which are marked in Figure 4-3A. Three out of four non-synonymous mutations in the *gacS*/*gacA* system occur in isolates from the same site, GW460, providing evidence for paired in situ evolution of this two-component system. The fourth isolate occurs in FW305, a spatially proximal site (Fig. 4-3B). In contrast, the two *rpoS*

SNPs occur in one isolate from FW301 and another from GW101 at the opposite end of the region, indicating more generalized selection processes. Overall, multiple non-synonymous mutations in the same gene is highly unlikely, but in the case of this probable pathway, all four genes show two non-synonymous mutations in two different isolates. This evidence raises *gacS*/*gacA* and *nlpD*/*rpoS* as critical genes for tuning transcriptional regulation in our sampling environment.

4.2.4 Evidence for gene loss in clonal isolates affects transcriptional regulators within wells

Mutation is one mechanism for evolution within these groundwater sites, but we also searched for traces of gene loss or gene gain via horizontal transfer within the aquifer wells. Within each *Pseudomonas* strain, we completed a pipeline to identify regions differentially present in otherwise closely related isolates. To accomplish this, the quality-filtered reads from each strain were mapped to each assembled contig within the subgroup, and any reads which mapped to some genomes but not others were flagged and grouped into common regions for analysis. Any region with greater than 100 reads mapped from at least one sample, and fewer than 10 from another, were exported as putative instances of gene loss/gain. We used the original contig annotations to contextualize these regions in terms of gene content and upstream or downstream impacts.

Within strain A, we identified one region which is an example of a multi-gene excision in two otherwise clonal isolates from a shared well (Fig. 4-4A). The two genes which had differential presence among these isolates included a GntR family transcriptional regulator as well as a glycosyltransferase. When we studied the regions impacted in representative draft assemblies, we found these two genes were adjacent to each other and shared an upstream promoter region (Fig. 4-4B). The reads which map to some draft genomes but not others appear to capture a gene loss event, since the two isolates from site FW306 missing these reads have truncated hypothetical genes in the same genomic context (Fig. 4-4C). It is also worth noting that this event occurred in an otherwise 0.3 Mbp contig with continuous, high read depth over the region. We not only found that one of the truncated genes included a GntR family transcriptional regulator, but the two downstream genes preserved in all isolates include a DNA binding transcriptional activator, *cpdR*, and a sensor kinase, *rpfC*, which are likely co-expressed with the non-truncated site.

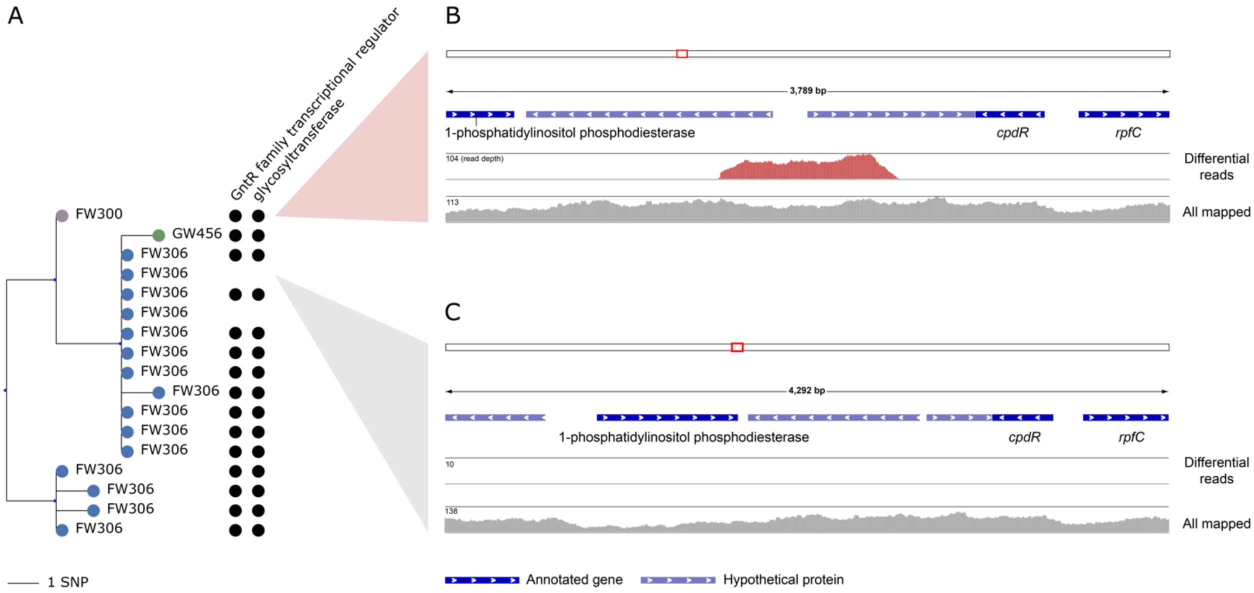

Figure 4-4. Differentially mapped reads within strain A. A) Genome tree of isolates from strain A (depicted in Figure 4-2), with leaves colored and named by the source well of each isolate. Next to the tree, black circles indicate the presence or absence of two genes, which happen to be adjacent when present in a genome. B) A representative example of a genome containing the two differentially mapped genes. The genome is assembled from FW300-N1A5 with reads mapped from sample FW306-02-F02-AA. Reads that map to this genome but not to the scaffold in (C) are labeled in red. C) A representative example of a genome lacking the two differentially mapped genes. The genome is assembled from FW306-2-11AB with reads mapped from the same sample as (B). In both (B) and (C), the top bar with a red box depicts the assembled contig and region of interest, respectively. Genes and predicted hypothetical proteins are labeled in blue. All mapped reads are shown in the lowest track in grey.

We found no indications of gene gain, consistent with the limited role horizontal transfer likely plays in the *Pseudomonas* genus; the genus is postulated to have a closed pan-genome (117).

Evidence for other instances of partial gene loss were identified in strains C and D, surprisingly impacting a periplasmic dipeptide transport protein (*dppA*) in both cases. In strain B, we found variability between isolates in the presence and absence of a cassette of arsenic regulatory genes including *aioA* and *acr3*, but it is unclear if this is due to horizontal transfer or simply incomplete genome recovery and assembly.

4.3 Discussion

The study of bacterial evolution and strain distribution within groundwater systems has received little attention despite critical roles in mitigating human impact and contamination. Here we identified a highly cultivable, site-relevant genus prevalent in the heavy metal and uranium contaminated Oak Ridge Watershed in TN. With diverse cultivation conditions, we recovered multiple isolates from 15 *Pseudomonas* strains, and each strain contained isolates from different sampling sites across the region. A stringent read mapping pipeline identified high confidence SNPs within each strain, and in one strain we found multiple non-synonymous mutations in a pathway likely under positive selection. There was also evidence for regulatory gene loss in otherwise clonal members of one strain, demonstrating active evolutionary processing within individual sampling sites.

Through strain-level isolate comparisons, we could record the recent evolutionary history of *Pseudomonas* strains recovered and cultivated from groundwater sites. To our knowledge, this work presents the largest grouping of whole genome isolate sequences sampling strains within an aquifer environment. Our high-throughput, nanoliter library preparations extended the boundaries of throughput for this type of sequencing effort. We also demonstrate for the first time how recovering and sequencing multiple isolates within a strain recovered from groundwater can reveal recent evolutionary changes occurring in situ.

Although ability to mine closely related genomes for recent adaptations is a powerful technique, it carries limitations relative to more common unbiased approaches. In this work, we haven't approached exhaustive sampling of *Pseudomonas* isolates at the site, and it may be possible to uncover additional evolutionary diversity by increasing the isolate sample size. Naturally this approach relies on strictly bacterial strains which are cultivable, restricting its current utility particularly in environments such as groundwater with high prevalence of candidate phyla. Finally, we quickly encountered limitations due to the huge amount of unannotated gene content, which is especially challenging in understudied and highly metabolically diverse groundwater communities. This challenge reduced our ability to infer mechanistic change from SNPs and limited our interpretation of the impacts of gene loss within strains.

This study in groundwater strain evolution generates many new research directions as well as a rich data source which could be further mined for biological insight. While we demonstrated that isolate whole genome sequencing can identify *in situ* evolution, our recovered SNPs should be further characterized in cloning or knockout experiments under different stress conditions. The *Pseudomonas gacS*/*gacA* system, in particular, has been shown to mitigate a number of stresses from the environment, and evidence indicates that broad members of this protein family may have physical contact-dependent signal transduction aiding in biofilm formation (118,119). In terms of gene gain and loss, it could be fruitful to search for horizontally acquired genes or plasmids which are linked to particular geographic sites, or occur in isolates from phylogenetically distant clades. Replicating this experimental design with more isolation conditions and a broader array of genera could certainly generate new insight into how individual cells and strains are coping with a nutrient limited, perturbed, and dynamic environment.

4.4 Materials and Methods

4.4.1 Sample collection

Groundwater samples were acquired between July 2010 and January 2016 from groundwater wells at the Oak Ridge Field Research Site in Tennessee, USA. Sampling methods were performed as previously published, which we summarize here (96). Either a peristaltic or bladder pump using low flow provided a collection flow, which was initially stabilized by flowing 2 to 20 liters of groundwater until pH, conductivity, and oxidation-reduction (redox) values stabilized.

Each collection included extensive physical and geochemical measurements at the time of sample extraction. At the wellhead we measured bulk water parameters such as pH, dissolved oxygen (DO), conductivity, and redox, using an In-Situ Troll 9500 system (In-Situ Inc., CO, USA). We also collected sulfide and ferrous iron [Fe(II)] groundwater concentrations with the U.S. EPA methylene blue method (Hach; EPA Method 8131) and the 1,10-phenanthroline method (Hach; EPA Method 8146), then analyzed these on site with a field spectrophotometer (Hach DR 2800). Samples were then preserved for further analysis with EPA-approved and/or standard methods described in (96).

4.4.2 Strain isolation and standard growth conditions

Strains used in this study were isolated from groundwater and sediment collected from the Oak Ridge Field Research Center, TN. In general, small 1-2 mls aliquots of different groundwater or sediment samples were grown on either rich media (Luria-Bertani, tryptic soy, R2A, Eugon, Winogradsky), basal medium (4.67 mM ammonium chloride, 30 mM sodium phosphate, with vitamins and minerals mixes as previously described (120)) or amended filtered groundwater under aerobic or anaerobic conditions at 25 or 30 ^{o}C in the dark. Positive growth was identified by increase in culture turbidity. After sequential transfers followed by streaking on agar plates, single colonies from clonal isolates were obtained. Individual colonies were picked, restreaked for purity tests, and regrown in liquid media. Overnight liquid cultures were used to extract DNA for 16S rDNA based identification. After identification, axenic cultures were grown to mid-log phase, amended with sterile glycerol (to a final concentration of 30%), flash frozen with liquid nitrogen, and stored at -80 ^{o}C.

The strains were revived from their glycerol stocks by streaking onto Luria-Bertani or R2A agar plates. Individual colonies developed at 30 ^{o}C over 48 hours, which were then inoculated into corresponding liquid media and grown at 30 ^{o}C for 48 hours. At that point, cell pellets were collected by centrifugation for DNA extraction.

4.4.3 Whole genome sequencing

Cultures were reconstituted and genomic DNA was extracted for downstream library prep. DNA extraction was completed with the Qiagen DNeasy kit (Qiagen, Venlo, NL) according to the manufacturer's instructions. All samples were eluted in Qiagen's AE buffer: 10 mM Tris-Cl, 0.5 mM EDTA, pH 9.0. Samples were stored at -20°C until randomized plating into a 384-well plate for automated library preparation. The isolated genomic DNA was normalized to 0.2 ng/uL in 10 mM Tris (pH 8.0), and libraries were prepared using the Illumina Nextera XT kit at 1/12th reaction size on a TTP Labtech Mosquito HV. Final libraries were cleaned with SPRI beads and pooled before sequencing on an Illumina NextSeq 500 with 150 bp paired-end reads.

4.4.4 Whole genome de novo assembly

Libraries were sequenced on an Illumina NextSeq producing 2x150 bp paired-end reads. Each sample contained 2,071,301 ± 409,888 reads, excluding one failed sample with < 2,000 reads. The program Cutadapt v1.12 was used to remove adapter sequences with parameters -a CTGTCTCTTAT -A CTGTCTCTTAT (121). We performed sliding window quality filtering with Trimmomatic v0.36 using parameters (-phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:20 MINLEN:50) (122). All genomes were assembled de novo using SPAdes v3.9.0 with the following options (-k 21,33,55,77 --careful) (123). Genome quality was validated with the program checkM v1.0.6 using the lineage_wf pipeline with default parameters (108), and draft genomes with contamination < 10% and completeness > 95% were maintained. 16S rRNA gene sequences were recovered with RNAmmer v1.2 (–S bac –m ssu) and taxonomically classified with SINTAX (usearch v9.2.64) against the RDP 16S rRNA gene training set v16 with species names and the following parameters (–strand both –sintax_cutoff 0.8) (124,125).

We completed initial characterization of genomes by extracting AMPHORA genes and preparing masked alignments (109). Translated gene sequences were identified with the AMPHORA2 script MarkerScanner.pl with parameters -Bacteria -DNA. Full-length marker protein sequences shared by all genomes were combined with custom scripts and then aligned with MUSCLE v3.8.31 using default parameters (126). Alignments were masked with Gblocks v0.91b and (–t=p –b4=5) (127). Remaining amino acids were concatenated into one representative alignment for each genome and a maximum likelihood tree was constructed with RAxML v8.2.4 (raxmlHPC –f a –m PROTCATLGF –p 12945 –x 23899 -# 100) (128).

4.4.5 *Pseudomonas* genus analysis

Based on the AMPHORA tree of genomes classified in the *Pseudomonas* genus, we identified closely related subgroups and performed alignment and SNP calling within these strains. First, for each strain we aligned the genomes of the isolates by selecting a random reference from the group and running Parsnp v1.2 with parameters (–r ! –c). From the resulting summary files we recovered and reported the core genome percentage of each strain alignment (129). SNP calling and genome tree construction was completed with custom MATLAB scripts as described in (5). Briefly, SNPs were called by mapping quality-filtered reads to a co-assembly crafted from all members of the strain subgroup using SPAdes v3.9.0 as described above. Gene annotations to contextualize recovered SNPs and assign loci were generated by Prokka v1.12 (130).

We identified instances of gene gain or loss by mapping the quality-filtered reads of each isolate in a strain subgroup against all other strains in the group. Read mapping was completed with bwa v0.7.5 using the bwa index and mem algorithms with default parameters (131). We used SAMtools to filter unmapped reads, then annotated the reads with BLAST v2.4.0+ (blastn with default parameters) against the NCBI non-redundant nucleotide collection (nt) as well as alignment to the annotated de novo assemblies (132,133). Reads differentially mapped between samples within a subgroup were quantified and summarized with custom scripts, then visualized along with Prokka annotations in the Integrative Genomics Viewer v2.3.94 (134).

Chapter 5 Conclusions

Microbial genomes show a rapid ability to exchange genes, evolve, bloom, and assemble in response to local environmental changes. I sought to track some of these dynamic qualities by developing new molecular biology techniques, with an aim to target individual microbial cells and also use materials and methods that would be accessible to the broader microbial ecology community. I extended this approach to study the spatial structuring of bacterial genomes, both at the microscale and across geographic space. The findings in this thesis, summarized below, highlight continuing opportunities in method development and genomic analysis that extend beyond 16S rRNA gene surveys and provide new classes of genomic information.

5.1 Single-cell capture and linked amplification via epicPCR identify species paired with target functional genes

In the second chapter of this thesis, I presented a novel technique to separate single bacterial cells directly from environmental communities, and link functional genes to phylogenetic indicators in a culture-independent design. This protocol combined previous emulsion-based research, and added a hydrogel encapsulation to enable microbial lysis. A series of control reactions showed perfect specificity for synthetic functional gene constructs, and demonstrated a broad profile of 16S rRNA gene recovery even without stringent chemical and enzymatic lysis. I concluded by performing a proof-of-principle in a lake water ecosystem, recovering the host species of dissimilatory sulfite reductase, variably distributed within the *Deltaproteobacteria*. This approach has enormous potential to spread within the academic community, since it requires no special equipment and focuses sequencing costs on a targeted hypothesis.

The concept of linking target genes within single bacterial genomes with massive throughput is inspirational and enabling, but as a new technique there is a large parameter space for possible improvement. The complexity of the multi-stage protocol is non-trivial, so further efforts to incorporate premixed solutions and reduce reagents would improve the workflow. Emulsion techniques are highly sensitive to initial cell loading, so one limitation is the continued need to produce accurate cell counts prior to sample processing. Also, while targeted primers can streamline sequencing costs, they generate bias and may miss some variants diverged at the priming sites; careful design with a large multi-gene alignment is necessary for new functional gene targets. Finally, as with many single-cell techniques, recovering strong data becomes more difficult for rare functional genes and hosts. Methods to physically sort out hydrogels carrying cells of interest, such as (41), would pair well with this approach and provide rich information for rare targets.

Future directions for targeted designs using hydrogel chemistry and highly parallel emulsions are numerous. On the technical side, the incorporation of simple, off-the-shelf droplet generators could dramatically improve the cell loading and PCR outcomes, moving closer toward quantitative readouts. Scientifically, we have early indications that linking common barcodes to multiple functional targets could provide rich data on linked pathways carried by single genomes. Another obvious extension would be recording the dynamic connections between bacteria and integrated phage or CRISPR arrays. Small adjustments could also be made to the hydrogel pore size in order to ensure capture of plasmids together with host genomes, allowing fusion constructs between antibiotic resistance cassettes and transient host species. Overall this method provides a powerful intermediate between 16S rRNA gene surveys and whole genome sequencing to provide functional information for targeted research questions.

5.2 Hydrogel capture enables spatial sequestration of bacterial aggregates for genomic analysis

In the third chapter, I presented a series of experiments expanding from the concept of connecting genes within genomes to connecting genes between genomes, in order to record microscale spatial structure. Two clade-targeted designs were used to assay spatial partners of the *Streptococcus* genus and the *TM7* phylum. These resulted in qualitative spatial partners that had some overlap between replicates, genus-level co-aggregation support in the literature, and differences recovered between sites. I also presented two studies which use droplet-specific barcodes to tag any available 16S rRNA genes in the same droplet. These studies show some reproducibility across samples, as well as some expected performance in fixed control cells, but also highlight ongoing challenges in control design and initial sample handling. While technically daunting, pushing forward this approach towards spatial sequencing has the potential to add a completely new dimension to microbial studies, analogous to the role Hi-C carried in mammalian genomic analysis (135).

The primary challenges remaining in microscale spatial sequencing center around control design and cell loading. Despite efforts to spike in freely suspended negative control cells as well as glutaraldehyde-fixed positive control aggregates, we observed inconsistent performance and poor amplification in the case of the fixed controls. Cell loading and sample handling remain challenging, particularly for biofilm studies; biofilm dispersal that preserves biological structure without high disruption is difficult to achieve via standard emulsification techniques. We suspect the high rate of background connections recovered in our barcoding studies represent large quantities of single cells pulled away from the bulk aggregates, which then randomly disperse and result in spurious connections. A future focus on sample handling or naturally suspended particles may reduce the background noise and provide a cleaner signal from the assay.

Based on the groundwork presented in this thesis, the most promising future direction for microbial spatial sequencing could be a transition towards a bulk hydrogel format, similar to efforts for single-cell sequencing and *in situ* RNA-Seq (20,87). This format has more technical parallels with combinatorial FISH designs and would simplify sample handling and minimize biofilm disruption. Controls would also become straightforward, since different cultures could be spatially arrayed across a 2D bulk hydrogel grid and amplified to test for cross-talk. With improved methodology, the ability to sequence the spatial structuring of bacteria within microenvironments has immediate applications in a variety of fields. There are still many open questions in oral microbiology, with poorly characterized candidate phyla in close proximity to putative host species. Improved spatial methods could even expand beyond bacteria, and provide a high-throughput view into predatory microeukaryotes in the environment or connections between immune cells and their targets.

5.3 Deep-branching *Pseudomonas* strains show recent regulatory adaptation within sampling sites

The fourth chapter presented covers a study of bacterial isolates gathered across and environmental gradient, revealing recent adaptive changes to a challenging environment. I generated close to 300 draft genome sequences from diverse isolates collected across a nuclear-contaminated watershed in Tennessee, using new instrumentation to complete library preparations in nanoliter volumes. Half of these genome sequences belong to the genus *Pseudomonas* and group into fifteen strains. SNP analysis within one of these strains showed adaptive selection impacting a two-component regulatory system, likely involved in iron uptake control. There were also signs of transcriptional regulatory gene loss occurring within sites. These results highlight the ability of strain-level whole genome sequencing to recover recent adaptations, even in dynamic and fast-flowing groundwater environments.

Key limitations of this study include the requirement for readily cultivated isolates as well as the restrictions of working with draft genomes. Although the *Pseudomonas* genus has a number of properties which make it interesting for study, a critical gating factor for our collection was the ease of isolation on a small number of test media and aeration combinations. Future studies would benefit from higher throughput testing of microbial media to produce a more diverse set of strains for study. From a computational perspective, I could only assemble draft genomes with approximately 200 contigs for each isolate. The nature of these data led to a restrictive SNP calling pipeline which likely eliminated some true positives, and also made it difficult to identify plasmids which may have a different profile of exchange and adaptation.

Despite these limitations, whole genome sequencing remains the gold standard for understanding microbial functional capacity within a complex microenvironment. The cost for microbial genome sequencing is decreasing every year, while new technologies including PacBio carry the promise of easily closed genomes in the near future (136). For the environmental site presented in this study, there are multiple directions for additional research. It would be useful demonstrate the adaptive fitness of identified isolates under metal or oxidative stress in vitro, and also to test the combinatorial impact of mutations accumulated in the two-component regulatory system within one strain. Since such rich information was extracted from a single genus and strain, follow-up studies on sediment-attached or anaerobic clades could reveal new routes of adaptation. Of course the most ambitious extension of this study would combine themes from Chapters 2 and 3 to generate single cell genomes in high-throughput, and efforts to generate this type of technology are actively in development (20,137–139).

5.4 New assays for expanded functional and spatial awareness in microbial communities

The unifying theme of this work features the functional and spatial plasticity of microbial genomes, which necessitates technology improvement for added insight into individual and community function. The past thirty years have generated an abundance of molecular biology tools and techniques that are rarely combined and often become underutilized. New molecular tools that could aid in novel technique development include tagmentation, split-and-pool techniques, and long construct sequencing (140). These join a host of specialized enzymes that can edit, append, and alter the direction of standard reactions. The beauty of molecular technology development is that it can always be scaled down for improved cost and throughput, so both academic and increasingly industrial efforts are pushing to miniaturize workflows.

These efforts join with increasing research extending the boundaries of microbial understanding, moving beyond surveys into functional insight. In addition, there are many opportunities to expand beyond bacteria and bridge between bacteria and their natural predators and prey in other domains of life. Future efforts should push towards developing and improving techniques that record the dimensions in which microbes reside and the dynamics of their gene flow and adaptation.

Appendix A epicPCR supplementary information

A.1 Supplementary Methods

A.1.1 epicPCR Reagents (in addition to solution reagents)

Ammonium persulfate (for molecular biology, ≥98.0%, Sigma, St. Louis, MO, USA)

TEMED (N,N,N′ ,N′ -Tetramethylethylenediamine, ≥99.5%, Sigma)

Diethyl ether (water-saturated, ≥99.5%, Sigma)

UltraPure DNase/RNase-Free Distilled Water (Life Technologies, Grand Island, NY, USA)

Ethyl acetate (water-saturated, ACS grade, ≥99.5%, BDH, Poole Dorset, UK)

Agencourt AMPure XP - PCR Purification (Beckman Coulter, Danvers, MA, USA)

Ethanol (200 proof, VWR, Radnor, PA, USA)

Ready-Lyse Lysozyme Solution (Epicentre, Madison, WI, USA)

Proteinase K from *Tritirachium album* (for molecular biology, Sigma)

BSA (molecular biology grade, NEB, Ipswich, MA, USA)

Tween 20 (for molecular biology, Sigma)

Deoxynucleotide (dNTP) Solution Mix (10 mM each, NEB)

Phusion Hot Start Flex DNA Polymerase (NEB)

Ethylenediaminetetraacetic acid (EDTA, suitable for cell culture, Sigma)

SYBR Green I Nucleic Acid Gel Stain (10,000X, Invitrogen, Waltham, MA, USA)

E-Gel 1 Kb Plus DNA Ladder (Invitrogen)

A.1.2 epicPCR Equipment

1.5 ml Safe-Lock Microcentrifuge Tubes, Polypropylene (Eppendorf, Hamburg, DE)

2 ml Safe-Lock Microcentrifuge Tubes, Polypropylene (round-bottom, Eppendorf)

PCR 8-Well Tube Strips with Individually Attached Caps (VWR)

Microcentrifuge (Microcentrifuge 5415D, Eppendorf)

Thermal-cycler (C1000 Touch Thermal Cycler, Bio-Rad, Hercules, CA, USA)

BD Falcon 35µm Cell Strainer in 12x75 mm Polystyrene Tube (Corning, Tewksbury, MA, USA)

2 mm glass beads (Andwin Scientific, Schaumburg, IL, USA)

DynaMag-2 Magnet (Life Technologies)

E-Gel iBase and E-Gel Safe Imager (Invitrogen)

E-Gel EX Agarose Gels, 1% (Invitrogen)

A.1.3 epicPCR Solutions

Acrylamide solution *(store at 4°C)*

12% Acrylamide (for molecular biology, ≥99.5%, Sigma)

0.32% BAC (N,N′ -Bis(acryloyl)cystamine, suitable for electrophoresis, Sigma)

1X TK buffer *(recommended filter through 0.2 µm, store at RT)*

20 mM Tris-HCl (pH 7.5, Teknova, Hollister, CA, USA)*

60 mM KCl (≥99.0%, VWR)*

*autoclave the two liquid stocks before combining

STT emulsion oil *(store at RT, should be prepared fresh every two weeks)*

4.5% Span 80 (Sigma)

0.4% Tween 80 (Sigma)

0.05% Triton X-100 (molecular biology grade, EMD Millipore, Billerica, MA, USA)

v/v in Mineral oil (light, suitable for cell culture, Sigma)

ABIL emulsion oil *(store at RT)*

4% ABIL EM 90, a surfactant (Evonik, Mobile, AL, USA)

0.05% Triton X-100 (molecular biology grade, EMD Millipore)

v/v in Mineral oil (light, suitable for cell culture, Sigma)

A.1.4 epicPCR Procedure

Polyacrylamide bead formation

To prepare polyacrylamide beads containing either cells or acrydited control molecules, we modified a polymerization protocol from (41). This involved the preparation of an aqueous suspension and then emulsification in an oil-surfactant solution. The 255 µl aqueous suspension included suspended cells or acrydited molecules, 0.98% ammonium persulfate (25 µl 10% APS), 9.4% acrylamide and 0.25% BAC (200 µl acrylamide solution). This suspension was applied to 600 µl STT emulsion oil, which was inverted and well-mixed before use, in a 2 ml round-bottom microcentrifuge tube and then vortexed for 30 s at 3000 rpm. We added TEMED to an aqueous concentration of 8.9% (25 µl TEMED) to catalyze the polymerization and vortexed for an additional 30 s at 3000 rpm, then let the emulsion polymerize for 90 min. Polyacrylamide beads were extracted with diethyl ether as described below, then filtered through a 35 µm cell strainer and transferred to a 1.5 ml microcentrifuge tube. Filtered polyacrylamide beads were stored at 4 °C and resuspended before subsequent lysis (described in main text).

Diethyl ether extraction for Span 80/Tween 80/Triton X-100 emulsions

When we phase-separated the emulsion oil from the polyacrylamide beads, we used an extraction protocol adapted from (52). We added 800 µl of diethyl ether (the upper layer of water-saturated mixture) to each round-bottom tube containing an emulsion, then immediately flicked and inverted the tubes to mix the emulsions with the ether in order to form a visible precipitate. The ether/oil mixture surrounding the precipitate was discarded and replaced with 1 ml nuclease-free water, followed by mixing and inversion of the tubes.

Samples were transferred to standard microcentrifuge tubes and centrifuged for 30 s at 12,000 *g*. We observed three layers form: a bottom layer of polyacrylamide beads, a middle cloudy layer of oil/water, and a top milky layer of oil. The top oil layer was removed and discarded without disturbing the lower layer of polyacrylamide beads, then additional nuclease-free water was added and polyacrylamide beads were resuspended by flicking and inversion. The centrifugation, oil removal, and wash steps were repeated until there was no remaining oil forming an upper phase (approximately five washes). After the final wash, all the water was removed from the beads and beads were resuspended in 1 ml 1X TK buffer.

Emulsion-concatenation library preparation

To form initial fusion products, we combined a PCR mix with polyacrylamide bead templates and added the suspension to ABIL emulsion oil, which is more thermostable than STT oil (52). The 100 µl PCR mix included 45 µl of polyacrylamide beads combined with PCR reagents and emulsion stabilizers (1X Phusion HF buffer, 1 mM $MgCl^2$, 250 µM each dNTP, 50 ng/µl BSA, 0.2% (v/v) Tween 20, and 0.16 U/µl Phusion Hot Start Flex). Additional primers and polyacrylamide beads used for specific samples are specified in Table A-3, with sequences in Table A-2. This mixture was placed in a 2 ml round-bottom microcentrifuge tube along with 900 µl ABIL emulsion oil. We also added four 2 mm autoclaved glass beads to the emulsion components in order to promote polyacrylamide bead separation during the emulsification process. The oil and aqueous phases were vortexed at 3000 rpm for 1 min, then aliquot into PCR tubes for thermocycling (94 °C 30 s; 33 cycles of 94 °C 5 s, 52 °C 30 s, 72 °C 30 s; 72 °C 5 min; 10 °C hold). Following amplification, the aliquots from each sample were pooled, supplemented with 1 mM EDTA, extracted with diethyl ether and purified with a modified AMPure XP protocol as described below.

Following the initial fusion reaction, we nested within the fusion products for increased specificity in the final library. We used a standard PCR mix (1X Phusion HF Buffer, 200 µM each dNTP, 0.02 U/µl Phusion Hot Start Flex) and prepared four replicate 25 µl reactions for each sample. The reagents were combined with nested primers (Table A-2, Table A-4), blocking primers (3.2 µM U519F_block10, 3.2 µM U519R_block10), and 2-5 µl of purified product from the previous fusion reaction. The thermocycling program (98 °C 30 s; 40 cycles of 98 °C 5 s, 52 °C 30 s, 72 °C 30 s; 72 °C 5 min; 10 °C hold) contained 40 PCR cycles by default. We reduced the number of cycles for the nested reaction whenever possible based on qPCR Ct values collected prior to the final nested reaction; these Ct values were collected using the same reaction conditions plus 0.5X SYBR Green I. Following amplification of the final nested reactions, the four replicate reactions were pooled and purified according to the modified AMPure XP protocol below.

The fused, nested products underwent a final, short amplification with Illumina adapters, then samples were pooled and submitted for sequencing. For each sample, we first assembled four replicate reactions using the standard Phusion Hot Start Flex reaction conditions. In the replicate reactions for a single sample we used 3.3 µM PE-PCR-F plus 3.3 µM PE-PCR-XXX to serve as a sample barcode (Table A-5). We amplified the libraries (98 °C 30 s; 7 cycles of 98 °C 30 s, 83 °C 30 s, 72 °C 30 s; 10 °C hold) and then pooled replicate reactions and purified with AMPure XP beads according to the modified protocol below. The appropriate amplicon size was confirmed on a 1% agarose E-Gel according to the manufacturer's instructions. Barcoded sample libraries were pooled in equal stoichiometric ratios and sequenced on an Illumina MiSeq with 20% phi-X spike-in to provide template diversity. We sequenced paired-end libraries with 250 bp reads in both directions and an 8 bp sample barcode read.

Diethyl ether extraction for ABIL EM 90/Triton X-100 emulsions

For the phase-separation of soluble fusion products from ABIL EM 90 oil emulsions, we again adapted a protocol from (52). Each sample was pooled and centrifuged at 13,000 *g* for 5 min at 25 °C. The upper (oil) phase was discarded and replaced with 1 ml diethyl ether (upper layer of water-saturated mixture), then vortexed to mix. Samples were centrifuged for 1 min at 13,000 *g* to separate the phases so that the upper phase could be discarded. This ether wash was repeated, then the same extraction was performed with ethyl acetate (upper layer of water-saturated mixture). We performed two more extractions with diethyl ether, then disposed of the upper phase. Samples were left open in a chemical hood for 10 min so the remaining diethyl ether could evaporate. For each sample we recovered 100-150 µl from the bottom phase into a fresh 1.5 ml microcentrifuge tube for purification prior to the nested PCR.

Modified AMPure XP purification

Our approach follows the manufacturer's protocol with the following variations. The AMPure XP beads were always equilibrated to room temperature (~30 min) before use. The beads were added to 1.5 ml microcentrifuge tubes in a ratio of 0.9 µl AMPure XP beads per 1 µl of PCR product. All mixing steps were completed by gentle vortexing or flicking. Upon addition of the AMPure XP beads, the solution was mixed and incubated for 13 min at room temperature. Two ethanol washes following magnetic separation were performed with 500 µl of 70% EtOH, and then the beads were air-dried for 15-20 min. The elution buffer (Buffer EB, Qiagen, Venlo, NL) was incubated with the beads for 7 min, then tubes were placed on a magnet for 2 min. The eluate was collected and transferred to a fresh tube.

A.1.5 epicPCR Accessory Procedures

Preparation of synthetic control polyacrylamide beads

In order to produce 348 bp segments of acrydited DNA sequence to incorporate into our positive and negative control polyacrylamide beads, we synthesized the sequences without the modification and then added the acrydite modification via PCR. The un-modified template DNA sequences (16S-V4neg and 16S-V4pos, Table A-1) were amplified in five replicate reactions using Phusion Hot Start Flex DNA Polymerase (NEB). The 50 µl reaction conditions were composed according to the manufacturer's protocol and included 0.5 µM each of 16S-synthF and 16S-synthR, along with 10 ng/reaction of un-modified template DNA (Table A-1). We cycled with standard conditions (98 °C

30 s; 25 cycles of 98 °C 5 s, 66 °C 30 s, 72 °C 30 s; 72 °C 10 min; 10 °C hold), then pooled the replicate reactions and purified with the MinElute PCR Purification Kit (Qiagen). The modified 16S-V4neg and 16S-V4pos sequences were used along with the independently synthesized dsrB-synth to attach to polyacrylamide control beads. The attachment was accomplished by mixing these acrydited amplicons with acrylamide solution and polymerizing as described in 'Polyacrylamide bead formation'.

Parallel epicPCR assay for rare target genes

In order to assay an increased number of cells and comprehensively sequence the species carrying *dsrB*, we performed the 21 m *dsrB*-16S rRNA gene fusion (abbreviated *dsrB*-16S) in multiple emulsion tubes and then combined the fusion products. Using previously polymerized polyacrylamide bead templates, we completed ten emulsion-concatenation reactions as described above and then combined the recovered aqueous phases. To purify and concentrate the fusion products, we used a MinElute PCR purification kit (Qiagen) instead of AMPure XP beads, concentrating ~1500 µl of recovered aqueous phase into a 10 µl final eluate. The concentrated fusion products were amplified with our nested PCR design for 40 thermal cycles, then labeled with a single sample barcode and flanked with Illumina paired-end sequencing adapters as described above. We loaded the final library on a 1% agarose E-Gel and excised the library band for purification using the Qiagen Gel Extraction Kit. Sequencing this library produced high-quality, paired-end fusion reads that matched our primer design and contributed to Figure 4.

Emulsion microscopy

In order to visualize emulsion droplets, both with and without polyacrylamide beads, we pipette dilute emulsions into a hemacytometer (Bright Line Counting Chamber, Hausser Scientific, Horsham, PA, USA). We combined 1 µl emulsion droplets with 9 µl mineral oil in a fresh microcentrifuge tube. This dilute emulsion was loaded into the hemacytometer and viewed at 100X resolution. We used the hemacytometer rulings to spot check the average droplet size of primary emulsions and also to quantify polyacrylamide bead loading in the secondary emulsion. In the secondary emulsion, out of nine 4000 μm^2 hemacytometer sections we observed 275 normal droplets and 4 droplets with two or more polyacrylamide beads. With our positive and negative spike-in ratios of approx. 2,000 control beads per 22,000,000 total beads, we expected and observed no negative fusion products owing to a 90% ratio of empty environmental beads. Fluorescence images presented in Figure A-4 were generated according to the SYBR Green I manufacturer's protocol.

dsrB primer design

We designed primers to target the beta subunit of dissimilatory sulfite reductase (*dsrB*) by using gene alignments and adapting primers from (54) and (55). Our dsrB-F1 primer (Table A-2) is equivalent to the dsr4R primer in (54). Our bridge primer, dsrB-R1_519R, contains a *dsrB* priming sequence based on the 1905 priming site in (55), but shifted over nine positions to fall at position 1896 of the *Desulfovibrio vulgaris* gene. It also contains added ambiguities: C→Y in position 6 of 1905 and C→ Y in position 9 of 1905. Finally, our i_dsrB-F3 nested primer is the reverse complement of 1929 with additional ambiguities Y→S in position 12 and Y→B in position 15 (55).

epicPCR sequence analysis and OTU clustering

For data analysis, we used the QIIME package with a few additional custom python scripts. To join the paired-end forward and reverse reads, we ran the QIIME command join-paired_ends.py with default parameters. The samples were demultiplexed and quality filtered with the QIIME command split_libraries_fastq.py (--min_per_read_length_fraction 0.40 -q 20 --max_barcode_errors 0 --max_bad_run_length 0). Following chimera identification using identify_chimeric_seqs.py (-m usearch61), we discarded chimeric sequences with a customized python script (version 2.7; https://github.com/sjspence/epicPCR/blob/master/discardChimeras.py). The remaining reads were filtered by length and expected fusion structure using custom python scripts. Our structure-filtering python scripts discarded any sequences that did not carry the expected forward, reverse, and bridge primers, then exported 121 bp of the captured 16S rRNA gene V4 variable region (version 2.7; https://github.com/sjspence/epicPCR/blob/master/filter*.py). If barcoded reads shared an identical droplet barcode and identical 16S rRNA gene sequence, we collapsed them into a single representative sequence using https://github.com/sjspence/epicPCR/blob/master/compressBar.py.

Our BLAST analysis for negative and positive control sequences relied on a simplified, custom BLAST database search. We used the synthetic designed sequences (16S_V4neg and 16S_V4pos, Table A-1) as a two-item database for our filtered 16S rRNA gene V4 fusion Illumina reads. The blastall 2.2.22 tool with default parameters identified reads with a significant match to our synthetic sequences (141). Both of these synthetic 16S rRNA gene V4 sequences were generated randomly and were thus highly divergent from any evolved 16S rRNA gene sequences.

For Operational Taxonomic Unit (OTU) assignment, we again relied on QIIME functions using default parameters unless otherwise specified. Starting from our stitched, quality-filtered, structure-filtered, length-trimmed 16S rRNA gene V4 sequences, we ran a series of commands to group and classify 97% identity sequence clusters. Our commands included pick_otus.py, pick_rep_set.py (-m most_abundant), assign_taxonomy.py, make_otu_table.py, and summarize_taxa.py. For datasets that compared multiple samples (e.g. Fig. 3), we rarefied the 16S rRNA gene V4 reads to the sample with the lowest read count using custom scripts in R. This rarefaction was performed after forming individual OTU tables but before summarizing taxonomic abundances. Computational commands are also presented step-by-step in a README file at https://github.com/sjspence/epicPCR/blob/master/README.md.

A.1.6 Sample collection, bulk 16S rRNA gene and *dsrB* gene library preparation

Sample collection

Water was collected from Upper Mystic Lake, (Winchester, MA, ~ 42 26.155N, 71 08. 961W) on Aug. 12, 2013 using a peristaltic pump and plastic Tygon tubing. Ethanol was applied to the end of the tubing and gloves were worn during collection to prevent contamination of samples during collection. A Hydrolab minisonde (Hach Hydromet, Loveland, CO, USA) attached to the end of the tubing recorded depth, dissolved oxygen, temperature, pH and specific conductance during deployment. Water from depth was allowed to flow through the tubing for 2 volumes (2 L) before

50 ml of water was filtered through a 0.22 μM filter in a 25 mm Swinnex-25 Filter Holder (Millipore, Darmstadt, DE) for DNA extraction. Filters were placed in a plastic bag and immediately placed on dry ice. For epicPCR, 7 ml of water was also added to 7 ml of 50% sterile glycerol in a 15 ml conical tube and immediately placed on dry ice. In parallel, aliquots were collected for nitrate and sulfate measurements via Ion Chromatography at the University of New Hampshire Water Resources Research Center. Blanks were collected by pumping 2 L of sterile water through the tubing before and after sampling to determine the influence of both contamination from the tubing and sampling method as well as carryover from the previous sample.

Bulk DNA extraction

Filters were stored at -80 °C until extraction. DNA was extracted from the filters using PowerWater DNA extraction kit (Mo Bio, Carlsbad, CA, USA) with an alternative lysis and proteinase K incubation step. Filters were removed from filter holders in a laminar flow hood and placed into the PowerWater Bead tube with a pair of sterile forceps as recommended. 1 ml of PW1 was added to the Bead tube, along with 20 μl of proteinase K (>600 mAU/ml, Qiagen). The alternative lysis protocol was followed by incubating samples at 65 °C for 10 minutes. Following the alternative lysis, the PowerWater protocol was followed, including horizontal vortexing with the recommended Mo Bio vortex adapter for 5 minutes and all subsequent steps. Purified DNA was stored at -20 °C.

Illumina 16S rRNA gene library preparation

The 16S rRNA gene libraries were prepared as previously described (106). Briefly, real-time PCRs were done first to normalize template concentrations and avoid cycling any templates past mid-log phase. PCRs for Illumina libraries were carried out as follows: 0.5 units of Phusion with 1X High Fidelity buffer, 200 μM of each dNTP, 0.3 μM of PE16S_V4_U515_F (5'-ACACGACGCTCTTCCGATCTYRYRGTGCCAGCMGCCGCGGTA

A-3') and PE16S_V4_E786_R (5'-CGGCATTCCTGCTGAACCGCTCTTCCGATCTGGACT

ACHVGGGTWTCTAAT-3') first step primers and approximately 40 ng of mixed DNA template were added for each 25 μl reaction. Additionally, 5X SYBR Green I nucleic acid stain (Molecular

Probes, Eugene, OR, USA) was added for real-time PCR. Samples were cycled with the following conditions: denaturation at 98 °C for 30 s, annealing at 52 °C for 30 s, and extension at 72 °C for 30 s. Samples were normalized to 20 cycles with the following dilution: $1.75^{(Ct-20)}$ or undiluted for samples with Ct larger than 20. The first step PCR was cycled as four 25 µl reactions for each sample with 20 cycles of amplification. PCRs were pooled and cleaned with Agencourt AMPure XP-PCR purification (Beckman Coulter) according to 'Modified AMPure XP purification' described above. Illumina-specific adaptors were added during a second step amplification, which include the sample specific barcode (index) sequences (Table A-5). The conditions for the second step PCR were similar to the first step, although 4 µl of the purified first step reaction was used as a template and 0.4 µM of each PE-PCR-F and the barcoded reverse primer was used with 9 cycles. Samples were cycled as four 25 µl reactions and cleaned with the Agencourt AMPure XP-PCR purification system using a modified protocol described above. Six samples (three samples, three blanks and three controls) were sequenced across 3 different MiSeq runs with multiple other samples.

Illumina dsrB gene library preparation

We amplified a region of the *dsrB* gene from bulk genomic DNA in order to compare the bulk *dsrB* diversity with epicPCR gene fusions. A 1:5 dilution of genomic DNA recovered from the bulk DNA extraction served as template for amplification with primers i_DSR1097AF (5'-CGGCATTCCTGCTGAACCGCTCTTCCGATCTGGAHTKGTGGATGGAAGA-3') and i_dsrB-F1 (5'-ACACGACGCTCTTCCGATCTYRYRGTGTAGCAGTTACCGCA-3'). We sourced the primer DSR1097AR from Giloteaux *et al.*, reverse complemented it (underlined), and added an Illumina adapter to produce i_DSR1097AF. From the same study we sourced DSR4R (underlined) and simply added an Illumina adapter to produce i_dsrB-F1 (55). For each of six samples, we prepared quadruplicate 25 µl PCRs (1X Phusion HF Buffer, 200 µM dNTPs, 0.5 µM i_DSR1097AF, 0.5 µM i_dsrB-F1, 0.5 U Phusion Hot Start Flex DNA Polymerase, 2 µl 1:5 genomic template). After cycling (94 °C 30 s; 25 cycles of 94 °C 5 s, 52 °C 30 s, 72 °C 30 s; 4 °C hold), quadruplicate reactions from each sample were pooled and purified with the Agencourt AMPure XP-PCR purification system using a modified protocol described above.

These six purified libraries were amplified again to add final Illumina adapters and barcodes.

Each sample was amplified in quadruplicate 25 µl PCRs (1X Phusion HF Buffer, 200 µM dNTPs, 0.4 µM PE-PCR-F, 0.4 µM PE-PCR-XXX, 0.5 U Phusion Hot Start Flex DNA Polymerase, 4 µl purified reaction from previous step). After cycling (98 °C 30 s; 15 cycles of 98 °C 10 s, 83 °C 30 s, 72 °C 60 s; 4 °C hold), quadruplicate reactions from each sample were pooled. Three samples were purified with the Agencourt AMPure XP-PCR purification system using a modified protocol described above. The other three were gel-purified to select for an 1,036 bp insert size which corresponds to the majority of published *dsrB* variants. The three non-gel purified and three gel-purified sample libraries were pooled in equal stoichiometric ratios and sequenced on an Illumina MiSeq with 20% phi-X spike-in to provide template diversity. We sequenced paired-end libraries with 250 bp reads in both directions and an 8 bp sample barcode read.

Bulk 16S rRNA gene sequence data processing

Paired end sequence data from each run was processed with SHERA (142), filtering out overlaps with less than 80% confidence (filterReads.pl with 0.8). Sequence and quality files were merged into fastq format with mothur make.fastq (143). Resulting fastq files were quality filtered and demultiplexed with QIIME split_libraries_fastq.py with the following options: truncate at positions in the read with quality scores less than 10 using ascii offset of 33 (-q 10 --max_bad_run_length 0 --phred_offset 33) and remove resulting reads shorter than 80% of the read length (-min_per_read_length .8). Primers were removed with a custom perl script, searching for the primer sequence 9 bp from either end of the forward and reverse position and allowing 4 bases of ambiguity at the end of the primer for mismatch repair. Processed reads were trimmed to 121 bp with a custom python script (version 2.7; https://github.com/sjspence/epicPCR/blob/master/filterLength.py) and then classified into OTUs according to ‘epicPCR sequence analysis and OTU clustering’.

A.2 Supplementary Figures

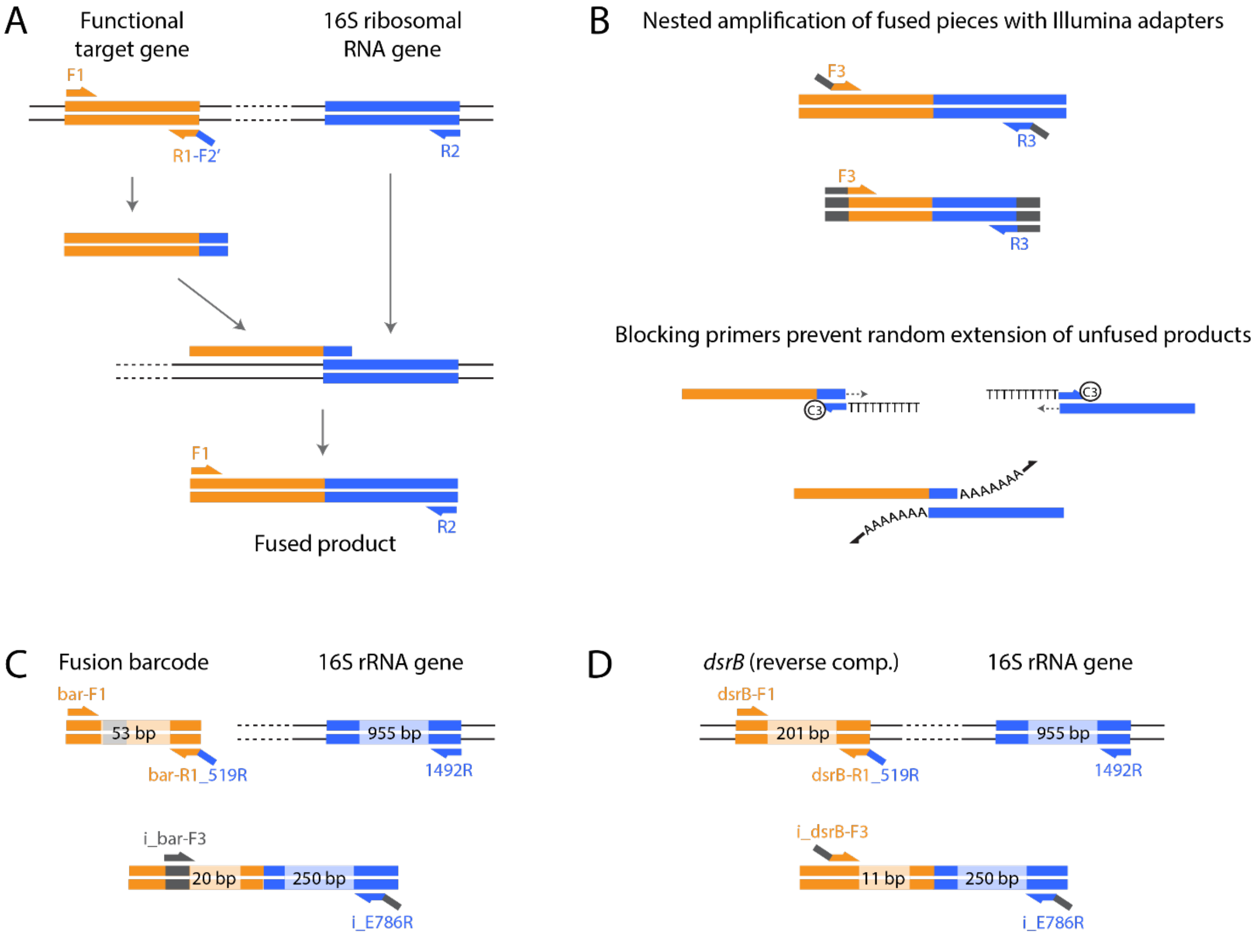

Figure A-1. epicPCR primers fuse target genes within droplets and then enrich for successful fusion constructs in a bulk nested reaction. A) Fusion PCR joins together two amplicons in a single reaction. The amplification first proceeds exponentially for the functional target gene from primers F1 and R1-F2' and linearly for the 16S ribosomal RNA gene from primer R2. Primer R1-F2' adds an overhang to the target gene amplicon that is specific to the start of 16S ribosomal RNA gene. Primers F1 and R2 are in excess over R1-F2', causing its depletion during the early cycles of PCR. After depletion of R1-F2', the 16S ribosomal RNA-specific overhang of the target gene amplicon primes the 16S ribosomal RNA gene creating a fused product. This fused product is subsequently exponentially amplified by F1 and R2. B) In the nested reaction, successful fusion products are amplified with Illumina adapters while partial fusion products are dampened by blocking primers. The blocking primers, added in excess, anneal to the universal 519R sequence but do not extend from the primer end due to a 3' 3-carbon-spacer. Instead, extension occurs from the 3' end of partial fusion products into the overhang region of the blocking primer, adding a string of A bases to the partially fused pieces. This A tail prevents partially fused pieces from annealing, extending, and generating spurious fusion products. C) Fusion construct design for fusions between a soluble molecular barcode and the 16S rRNA gene. The first row shows the initial fusion design and the second row shows the nested reaction design. D) Fusion construct design for fusions between bacterial *dsrB* and the 16S rRNA gene.

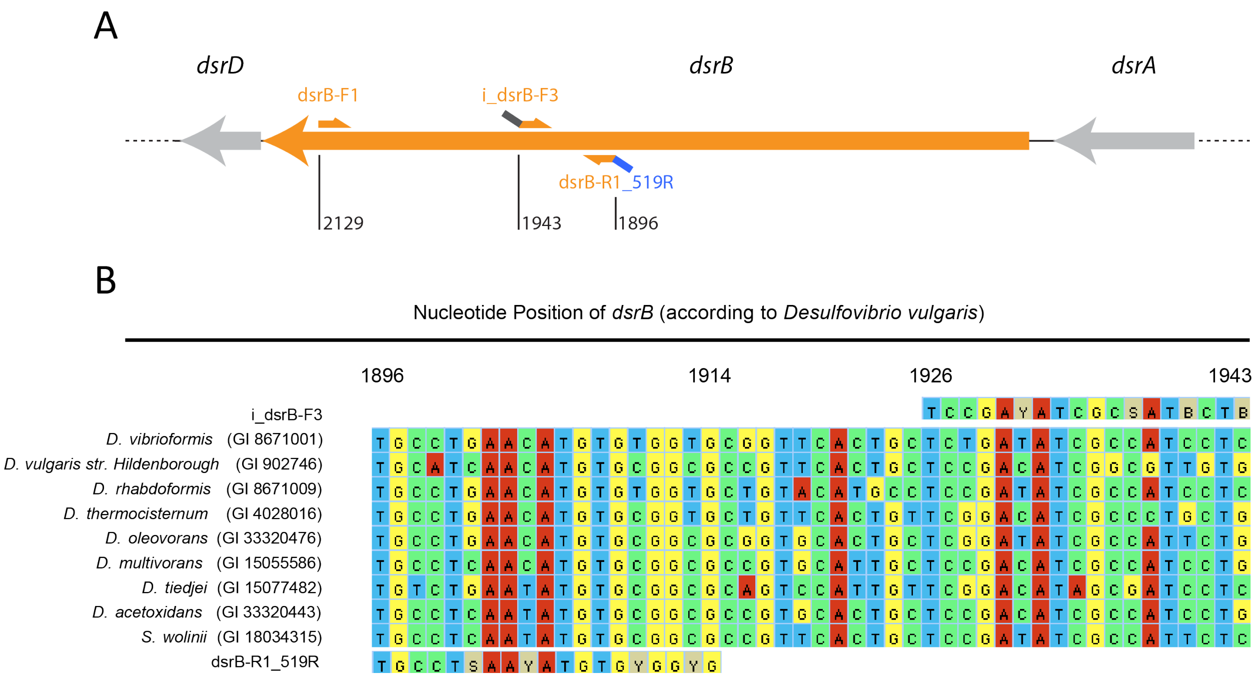
A
dsrD
dsrB
dsrA
dsrB-F1
i_dsrB-F3
dsrB-R1_519R
2129
1943
1896
B
Nucleotide Position of dsrB (according to Desulfovibrio vulgaris)
1896
1914
1926
1943
i_dsrB-F3 TCCGAYATCGCSATBCTB
D. vibrioformis (GI 8671001) TGCCTGAACATGTGTGGTGCGGTTCACTGCTCTGATATCGCCATCCTC
D. vulgaris str. Hildenborough (GI 902746) TGCATCAACATGTGCGGCGCCGTTCACTGCTCCGACATCGGCGTTGTG
D. rhabdoformis (GI 8671009) TGCCTGAACATGTGTGGTGCTGTACATGCCTCCGATATCGCCATCCTC
D. thermocisternum (GI 4028016) TGCCTGAACATGTGCGGTGCTGTTCACTGTTCGGACATCGCCCTGCTG
D. oleovorans (GI 33320476) TGCCTGAACATGTGCGGCGCGGTGCACTGCTCGGATATCGCCATTCTG
D. multivorans (GI 15055586) TGCCTCAACATGTGCGGCGCCGTGCATTGCTCCGACATCGCCATCCTG
D. tiedjei (GI 15077482) TGTCTGAATATGTGCGGCGCAGTCCATTGTTCGGACATAGCGATCCTC
D. acetoxidans (GI 33320443) TGCCTCAATATGTGCGGCGCCGTGCACTGCTCCGACATCGCCATCCTG
S. wolinii (GI 18034315) TGCCTCAATATGTGCGGCGCCGTTCACTGCTCCGATATCGCCATTCTC
dsrB-R1_519R TGCCTSAAYATGTGYGGYG

Figure A-2. Degenerate primers target the *dsrB* gene for epicPCR. A) A schematic showing the three *dsrB* primers in their approximate genomic context. Nucleotide positions below the primers are based on the *Desulfovibrio vulgaris dsrB*. B) A selection of nucleotide alignments demonstrating the genomic context and selected degeneracies for primers i_dsrB-F3 and dsrB-R1_519R (55).

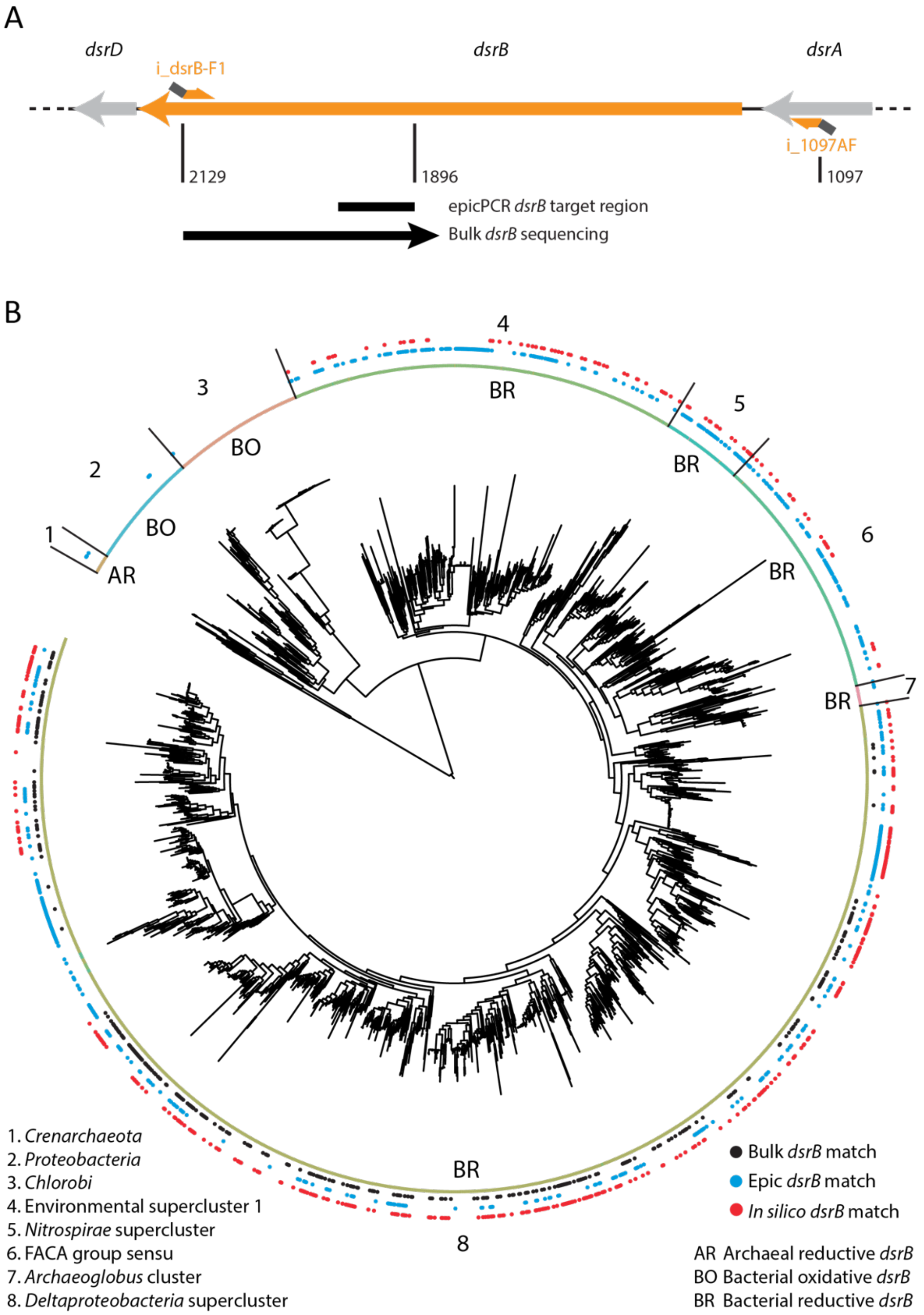

Figure A-3. Bulk dsrB gene fragment short-read sequencing provides a background distribution for observed epicPCR dsrB fragments. A) A schematic showing the bulk dsrB primers, modified from (55), in their approximate genomic context. Nucleotide positions below the primers are based on the Desulfovibrio vulgaris dsrB. Grey primer overhangs indicate Illumina adapter sequences. B) Distribution of bulk dsrB sequencing reads, epicPCR dsrB reads and in silico epicPCR dsrB matches in a tree of known dsrAB genes (46).

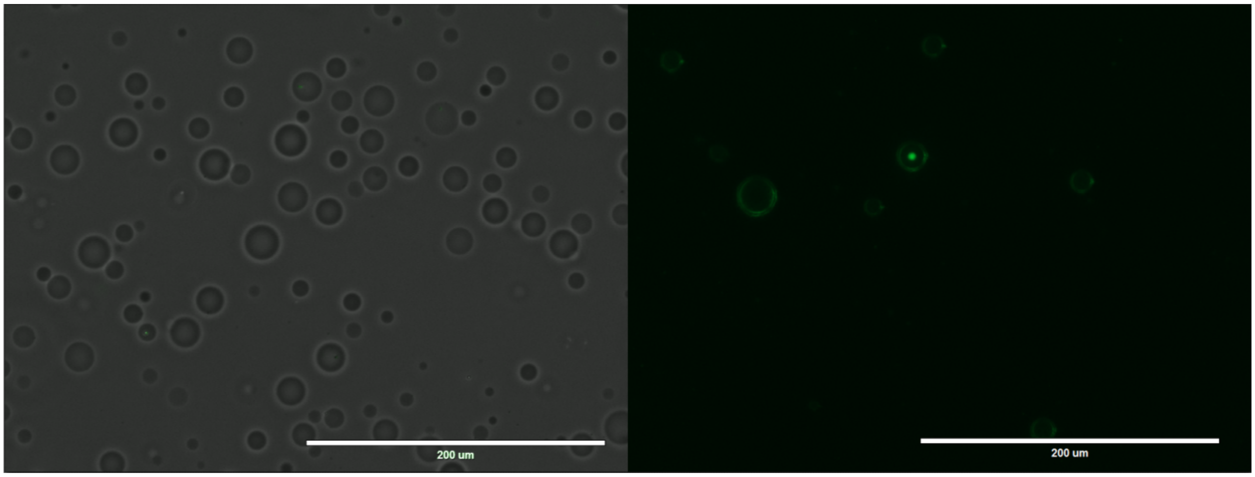

Figure A-4. Vortex-generated emulsions separate single cells or single polyacrylamide beads into nanoliter volume droplets. A) Single cells disperse into individual droplets, with the majority of droplets empty. This merged image shows bright-field emulsion droplets overlaid with a fluorescence image of SYBR-stained bacterial cells. B) Polyacrylamide beads in the secondary emulsion carry bacterial chromosomes as templates for fusion PCR. This fluorescence image shows a SYBR-stained bacterial genome in a polyacrylamide bead, suspended in emulsion oil prior to fusion PCR.

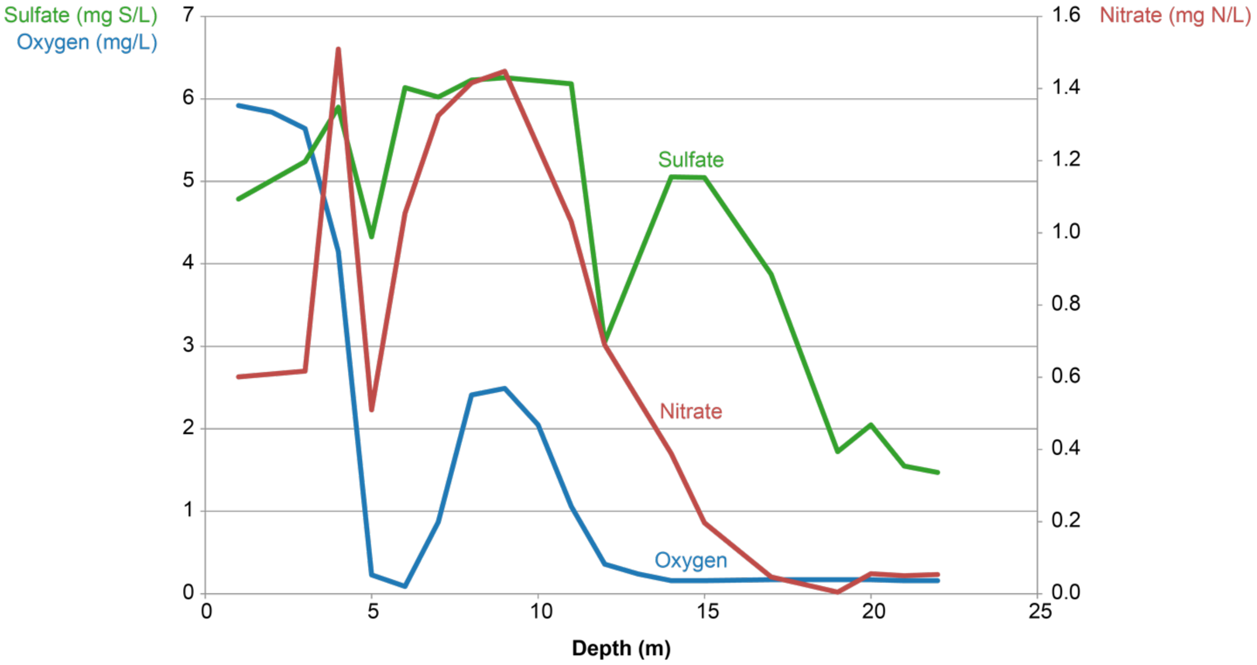

Figure A-5. Observed geochemistry at different lake depths collected on 8/12/2013. At a 2 m depth, oxygen predominates. At a 21 m depth, both oxygen and nitrate are depleted, but sulfate is still available as an electron acceptor.

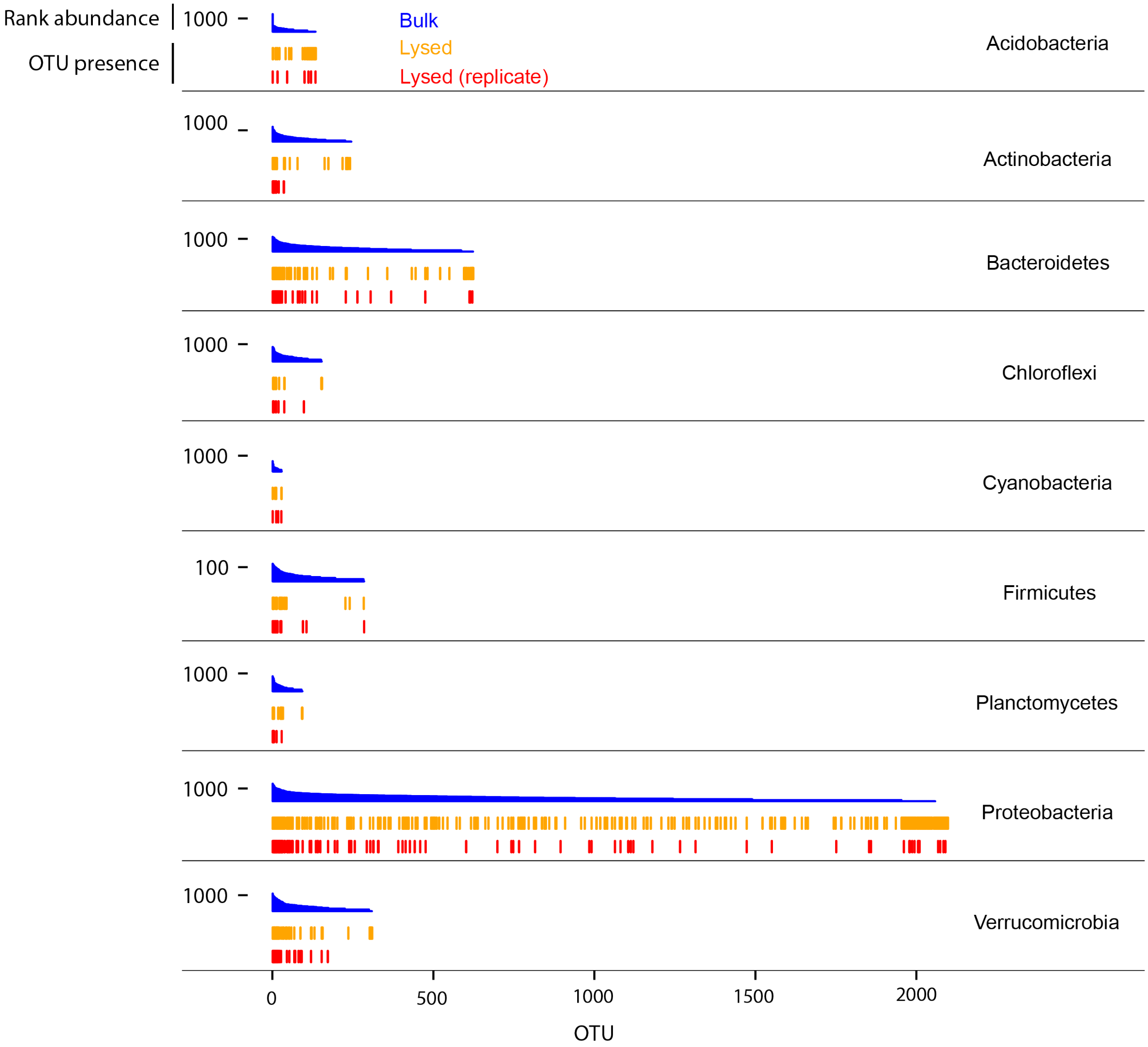

Figure A-6. Duplicate epicPCR barcoded libraries from the 21 m lake depth. OTUs are listed by phyla according to the rank-ordered abundance from bulk 16S rRNA gene sequencing. Below the bulk sequencing, the presence of an OTU in duplicate, lysed 21 m epicPCR libraries is indicated by an orange or red bar.

A.3 Supplementary Tables

Table A-1. Primers used for synthetic bead preparation. DNA was incorporated into polyacrylamide hydrogels via an acridite modification at the 5' end of the sequence (/5Acryd/). The synthetic *dsrB* sequence was synthesized directly with the acrydite attachment (dsrB-synth). The acrydite attachment was added to the synthetic 16S rRNA gene V4 sequences (16S-V4neg, 16S-V4pos) using an acrydited forward primer (16S-synthF) in a PCR (see Supplementary Methods). Colors indicate identical or reverse complement primers and corresponding priming sites.

Sequence Name	Sequence (5' →3')
dsrB-synth	/5Acryd/GTGTAGCAGTTACCGCAGAGGATGGCGATATCGGAGCATTGCACCACACATGTTCAGGCA
16S-synthF	/5Acryd/TCGAGGCCGTTCGTTAATTC
16S-synthR	GGAGCGTCCGGTATTGATTG
16S-V4neg	TCGAGGCCGTTCGTTAATTCCAGCAGCCGCGGTAATACGTAAACTACGATGGCACCAACTCAATCGCAGCTCGTGCGCCCTGAATAACGTACTCATCTCAACTGATTCTCGGCAATCTACGGAGCGACTTGATTATCAACAGCTGTCTAGCAGTTCTAATCTTTTGCCAACATCGTAATAGCCTCCAAGAGATTGATCATACCTATCGGCACAGAAGTGACACGACGCCGATGGGTAGCGGACTTTTGGTCAACCACAATTCCCCAGGGGACAGGTCCTGCGGTGCGCATTAGATACCCTGGTAGTCCAAAGTCGTAACAAGGTAACCCAATCAATACCGGACGCTCC
16S-V4pos	TCGAGGCCGTTCGTTAATTCCAGCAGCCGCGGTAATACATAGCCGCGCTATCCGACAATCTCCAAATTATAACATACCGTTCCATGAAGGCCAGAATTACTTACCGGCCCTTTCCATGCGTGCGCCATACCCCCCCACTCCCCCGCTTATCCGTCCGAGGGGAGAGTGTGCGATCCTCCGTTAAGATATTCTTACGTATGACGTAGCTATGTATTTTGCAGAGGTAGCGAACGCGTTGAACACTTCACAGATGGTGGGGATTCGGGCAAAGGGCGTATAATTGGGGACATTAGATACCCTGGTAGTCCAAAGTCGTAACAAGGTAACCCAATCAATACCGGACGCTCC

Table A-2. Primers used for epicPCR. We synthesized a fusion barcode with 20 degenerate bases that we spiked in at low concentrations in order to fuse it to any 16S rRNA genes available in each droplet. We performed barcode-16S rRNA gene fusions with bar-F1, 1492R, and bar-R1_519R as a bridge primer. We performed *dsrB*-16S fusions with dsrB-F1, 1492R, and dsrB-R1_519R as a bridge primer. Universal primer segments 1492R and 519R were drawn from (56) and *dsrB* primers were adapted from (54,55). Colors indicate identical or reverse complement sequences between Tables A-2, A-4, A-5.

Primer Name	Sequence (5' →3')
fusionBarcode	CGGCACAATCTCGTCGCGTCGACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNNNNNGATCATGACCCATTTGGAGAAGATG
bar-F1	CGGCACAATCTCGTCGCGTCG
1492R	GGTTACCTTGTTACGACTT
bar-R1_519R	**GWATTACCGCGGCKGCTG**CATCTTCTCCAAATGGGTCATGATC
dsrB-F1	GTGTAGCAGTTACCGCA
dsrB-R1_519R	**GWATTACCGCGGCKGCTG**TGCCTSAAYATGTGYGGYG

Table A-3. Samples, conditions, and primer sets used to produce particular epicPCR libraries. At the 2 m depth, 7 × 10^7 cells were suspended in polyacrylamide beads. At the 21 m depth, 1.4 × 10^7 cells were suspended in polyacrylamide beads. Use of lysis reagents is described in the main text. When control beads were spiked in, we added 0.5 µl 200X dilution of the initial bead preparation as described in Supplementary Methods. Concentrations of listed primers in the final fusion reactions were 100 fM fusion barcode, 1 µM F1, 1 µM R2, and 10 nM R1-F2'. Concentrations of listed primers in the subsequent nested reactions were 0.3 µM Nested F3 and 0.3 µM Nested R3.

Depth	Lysis reagents	Control beads	Fusion barcode	F1	R2	R1-F2'	Nested F3	Nested R3
2 m	-	-	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
2 m	+	-	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
2 m	+	+	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
21 m	-	-	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
21 m	+	-	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
21 m	+	+	+	bar-F1	1492R	bar-R1_519R	i_bar-F3	i_E786R
2 m	+	-	-	dsrB-F1	1492R	dsrB-R1_519R	i_dsrB-F3	i_E786R
2 m	+	+	-	dsrB-F1	1492R	dsrB-R1_519R	i_dsrB-F3	i_E786R
21 m	+	-	-	dsrB-F1	1492R	dsrB-R1_519R	i_dsrB-F3	i_E786R
21 m	+	+	-	dsrB-F1	1492R	dsrB-R1_519R	i_dsrB-F3	i_E786R
21 m	+	-	-	dsrB-F1	1492R	dsrB-R1_519R	i_dsrB-F3	i_E786R

Table A-4. Primers used for the nested PCR. Either the i_bar-F3 or i_dsrB-F3 primers were used in the forward direction, paired with i_E786R in the reverse direction. The blue and red segments are overhangs used for Illumina adapter addition (see Table A-5). The underlined segment of i_dsrB-F3 indicates a small degenerate sequence that was added to increase the sequence complexity of the amplicon library for improved Illumina image analysis. The blocking primers, U519R-block10 and U519F-block10, carry a 3-carbon spacer to prevent 3' extension; this forces the addition of A bases to the 3' end of any unfused pieces. Universal primer segments E786R and 519R/F were drawn from (56) and *dsrB* primers were adapted from (53). Colors indicate identical or reverse complement sequences between Tables A-2, A-4, A-5.

Primer Name	Sequence (5' →3')
i_bar-F3	ACACGACGCTCTTCCGATCT
i_dsrB-F3	ACACGACGCTCTTCCGATCTYRYRVAGVATSGCGATRTCGGA
i_E786R	CGGCATTCCTGCTGAACCGCTCTTCCGATCTGGACTACHVGGGTWTCTAAT
U519R-block10	TTTTTTTTTTGWATTACCGCGGCKGCTG/3SpC3/
U519F-block10	TTTTTTTTTTCAGCMGCCGCGGTAATWC/3SpC3/

Table A-5. Primers used for Illumina library preparation**.** The forward primer PE-PCR-F can pair with any of the reverse primers (PE-PCR-XXX). The numbered primer names indicate reverse primers with different Illumina barcode sequences that can serve as sample identifiers in pooled sequencing runs. The underlined sequence indicates the unique sample barcode within these reverse primers. Colors indicate identical or reverse complement sequences between Tables A-2, A-4, A-5.

Primer Name	Sequence (5' →3')
PE-PCR-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE-PCR-001	CAAGCAGAAGACGGCATACGAGATTCCGTGCGCCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-002	CAAGCAGAAGACGGCATACGAGATTGTTTCCCACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-003	CAAGCAGAAGACGGCATACGAGATGGTAATGAACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-004	CAAGCAGAAGACGGCATACGAGATGAAACTGGGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-005	CAAGCAGAAGACGGCATACGAGATACGGGCTGACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-006	CAAGCAGAAGACGGCATACGAGATATGAAGTATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-007	CAAGCAGAAGACGGCATACGAGATACTTATTGTCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-008	CAAGCAGAAGACGGCATACGAGATGGCGGGAAACGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-009	CAAGCAGAAGACGGCATACGAGATACACCTCGGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-PCR-010	CAAGCAGAAGACGGCATACGAGATCTCATTGGGCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Table A-6. 16S rRNA gene taxonomy recovered from the *dsrB*-16S fusion libraries. OTUs were assigned by grouping the 16S rRNA gene sequences into 97% identity clusters. The taxonomy was determined by Qiime based on the Greengenes database. The number of known and novel OTUs and reads recovered is indicated adjacent to taxonomic designations.

Phylogenetic level				*dsrB* OTUs		*dsrB* reads	
Class	Order	Family	Genus	Non-novel	Novel	Non-novel	Novel
Betaproteobacteria	Unclassified	Unclassified	Unclassified	0	1	0	1
Deltaproteobacteria	Unclassified	Unclassified	Unclassified	1	1	3	8
	Desulfarculales	Desulfarculaceae	Unclassified	0	1	0	2
	Desulfobacterales	Desulfobacteraceae	Unclassified	7	6	65689	318139
			Desulfococcus	0	2	0	13
	Desulfuromonadales	Desulfuromonadaceae	Unclassified	1	1	1	3
	Syntrophobacterales	Syntrophaceae	Unclassified	1	0	2	0
			Desulfomonile	9	8	642650	107
		Syntrophobacteraceae	Unclassified	5	3	1000489	1088
			Syntrophobacter	1	1	1	4
Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	0	1	0	1
	Oceanospirillales	Halomonadaceae	Halomonas	0	6	0	4086

Appendix B Spatial PCR supplementary information

B.1 Supplementary Methods

B.1.1 Baited cultivation panel designed for *SR1* co-cultivation

We recovered saliva from a subject with a high relative abundance of the *SR1* phylum, and used a panel of *Streptococcus* species to attempt co-cultivation of *SR1*. Strains of *S. intermedius*, *S. parasanguinus*, *S. mitis*, *S. tigurinis*, *S. australis*, and *F. nucleatum* (gram - control) were grown in RPMI to similar log-phase cell densities. We performed a light centrifugation (1,000 *g* for 5 min) on the 7 ml saliva sample to collect mammalian cells and mucins toward the bottom of the tube. The supernatant was transferred into a vacuum system for filtration, and then the pass-through was centrifuged again. We added 1,200 µl 1X PBS to resuspend the final pellet, then combined 2:2:1 ratios of RPMI media, cultured *Streptococcus* strains, and resuspended salivary microbes into two duplicate culture plates. One was grown in an anaerobic chamber, and the other in a microaerophilic chamber. Every three days the cultures were transferred into fresh media in another well.

After three days and seven days of growth, all combination cultures were checked for *SR1* presence with an *SR1*-targeted PCR. We used the GoTq Green Master Mix (Promega, Madison, WI, USA) according to the manufacturer's instructions, and added 2 mM additional $MgCl_2$. For primers we added 600 nM each SR1_183F (5'-ACGATGGTGAAATTCCGATG-3') and SR1_299R (5'-ATCGCGACCGGACATCAT-3'). The cultured cells were added directly in a 1/25 ratio and cycled for 95 °C 5min, 30 cycles of (95 °C 30 sec, 55 °C 30 sec, 72 °C 30 sec), 4 °C hold. We checked for visualization of the *SR1* amplified band on a 1% agarose gel, using high *SR1* concentration extracted gDNA as a PCR positive control.

B.1.2 Preparation of glutaraldehyde-fixed synthetic cell aggregates

Laboratory strains of E. coli and B. subtilis were cultivated and fixed in high concentration pools to prepare a positive control spike-in for cell-cell association. An *E. coli* K12 WT strain with an integrated chloramphenicol resistance cassette was cultivated at 37 °C in LB with 50% head space and 200 rpm shaking. A *B. subtilis* strain (AG174, trp- phe-) was also cultivated in LB with 90% headspace under the same temperature and shaking conditions (144). After diluting the cultures and allowing log growth, we used OD600 to estimate cell count for different ratios of cell combinations. While waiting for cell growth, we prepared glutaraldehyde solution (3% v/v) in 1X phosphate-buffered saline (PBS) (Corning, Corning, NY, USA).

We tested a variety of cell concentrations and fixation conditions for efficient aggregate formation, then stored treated cells to use as spike-in positive control aggregates in barcode epicPCR assays. We combined 500 million cells from each strain, pellet the cells at 600 *g* for 5 min, then resuspended in 100 µl PBS. Immediately, 100 µl 3% glutaraldehyde was added for a final concentration of 1.5%. The sample was mixed gently, then stored at room temperature for 2 hours. After fixation, cells were centrifuged at 8500 *g* for 1 min and the supernatant was discarded. We added 1 ml PBS and resuspended the cells by gentle inversion. The PBS wash was repeated, and then cells were resuspended in 100 µl PBS and 100 µl EtOH for storage at -20 °C until downstream use.

In our final replicated set of barcoding experiments, we spiked in the equivalent of 1 million cells each of *E. coli* and *B. subtilis*, fixed together in aggregate, together with the previously described *S. oneidensis* spike-in. We also prepared replicate libraries at each shear force level which included only cells from our positive and negative control strains. These each carried the equivalent of 4 million cells from each strain, e.g. 4 million *S. oneidensis* cells, plus the glutaraldehyde-fixed combination of 4 million *E. coli* cells plus 4 million *B. subtilis* cells taken from our ethanol stock.

B.1.3 Bulk 16S rRNA gene library preparation

For the clade-targeted assays, we thawed the 25% glycerol stocks of remaining oral collections by hand, and completed DNA extraction and 16S rRNA gene sequencing. For the DNA extraction, we recovered half the volume of each glycerol stock and placed the remainder at -80 °C. The recovered volume was centrifuged at 13,000 g for 1 min, the supernatant was discarded, and 100 ul 1X PBS was added. Each sample was gently resuspended with aspiration and low level vortexing, then transferred into the PowerSoil DNA Isolation Kit (Mo Bio, Carlsbad, CA, USA) for extraction according to the manufacturer's instructions. Recovered gDNA was quantified and stored at -20 °C in multiple aliquots.

In order to sequence the 16S rRNA gene V1/V3 variable region, we completed a two-step PCR protocol to amplify the product and add Illumina adapters. Three independent library preps were performed for each aliquot of source gDNA. Samples were normalized by completing duplicate qPCRs with 1:20 and 1:200 dilutions of gDNA, then using Ct values to dilute to a common input concentration. The qPCRs combined 280 nM PE-16S-V1V3-F and 280 nM PE-16S-V1V3-R (Table B-3) into a reaction with 0.5X SYBR Green I nucleic acid gel stain (Sigma-Aldrich, St. Louis, MO) and the standard Phusion High-Fidelity PCR Kit (New England BioLabs, Ipswich, MA) reagents according the the manufacturer's instructions. Reactions underwent cycling according to the program: 98°C 30 sec; 30 cycles of 98°C 30 sec, 52°C 30 sec, 72°C 30 sec; 4°C hold. Following qPCR normalization and cycle calculation, quadruplicate PCRs were performed under the same conditions minus the SYBR Green I. Quadruplicate reactions were then pooled and purified with Agencourt AMPure XP Beads (Beckman Coulter, Brea, CA) according to the manufacturer's instructions. One fourth of the final elution volume served as a template for a second step PCR.

A second step PCR was performed to add complete Illumina adapter sequences to the 16S rRNA gene amplicons. This reaction included 420 nM each of indexing primers PE-III-PCR-F and PE-IV-PCR-R (Table B-3). The primers were arrayed row- and column-wise to produce uniquely barcoded samples, and amplified with the Phusion High-Fidelity PCR Kit according to manufacturer's instructions. The thermocycling program included the following steps: 98°C 30 sec; 7 cycles of 98°C 30 sec, 83°C 30 sec, 72°C 30 sec; 4°C hold. We purified indexed samples with Agencourt AMPure XP Beads according to the manufacturer's instructions and quantified the final libraries with SYBR Green I and a standard curve. Libraries were combined in equimolar ratios and sequenced on an Illumina MiSeq with 2x250 bp paired-end reads.

B.1.4 Hydrogel-encapsulated plaque lysis conditions

Following hydrogel encapsulation, plaque samples were lysed with a combination of lysozyme, proteinase K, detergent, and heat treatment. A final concentration of 50 U/µl of Ready-Lyse Lysozme (Epicentre, Madison, WI, USA) was added and samples were gently mixed and incubated at room temperature overnight. Following incubation, beads were centrifuged at 12,000 *g* for 30 sec, then one third of the total volume was discarded and replaced with 1x TK buffer (see Section A.1.3). We added 110 ng/µl proteinase K from *Tritirachium album* (for molecular biology, Sigma, St. Louis, MO, USA) and 0.4% (v/v) Triton X-100 (molecular biology grade, EMD Millipore, Billerica, MA, USA), then mixed thoroughly. Samples were incubated at 37 °C for 30 min, then the proteinase K was digested during an incubation at 95 °C for 10 min. Three washes were performed by centrifuging samples at 12,000 *g* for 30 sec and replacing half the volume with fresh 1X TK buffer. Lysed hydrogels were filtered through a 100 µm cell strainer (Falcon, Nylon, Sterile, Corning), and the flow-through was transferred to a microcentrifuge tube for storage at 4 °C.

B.1.5 Final Illumina library adapter addition and sample barcoding

After nested amplification, approximately half of the eluted volume was used as template for a final amplification and adapter addition reaction. The reaction consisted of 0.02 U/µl Phusion Hot Start Flex DNA Polymerase (NEB, Ipswich, MA, USA), 1X HF Buffer, and 200 µM each dNTP, combined with 400 nM each of PE-III-PCR-F_fusion and PE-IV-PCR-R_fusion (Table B-3). Quadruplicate reactions were prepared for each sample, then pooled after thermocycling under the following program: 98 °C 30 sec, 7 cycles of (98 °C 30 sec, 83 °C 30 sec, 72 °C 30 sec), 4 °C hold. Samples were purified with the Agencourt AMPure XP - PCR Purification kit (Beckman Coulter, Danvers, MA, USA) according to the manufacturer's instructions.

B.1.6 Supplementary computational methods

For both clade-targeted and barcoding approaches, we used similar clustering, noise-filtering, and taxonomic assignment pipelines. In order to computationally optimize the processing of our large barcoded data set, we performed a dereplication step and then filtered out likely PCR chimeras and artefacts. After primer filtering, we grouped the data into unique representative sequences (zero-radius OTUs) with a custom script. Then the usearch v9.2 unoise2 algorithm discarded likely chimeras, phiX sequence, and low complexity DNA (–minampsize 3) (145). Our clade-targeted pipelines used the usearch v6.1 chimera filtering algorithm. In order to select representative sequences for each OTU, we used sequence-based and taxonomic clustering. Our clade-targeted segments were clustered at 97% sequence identity with usearch v8, then assigned taxonomy with the QIIME implementation of the Mothur naïve bayes classifier (58,143). Barcoded reads were assigned taxonomy via the analogous SINTAX classifier from usearch v9.2, and then reads matched to the same taxonomic level (probability > 0.8) were grouped into a representative 'taxonomic' OTUs for analysis (–strand plus –sintax_cutoff 0.8) (124).

B.2 Supplementary Figures

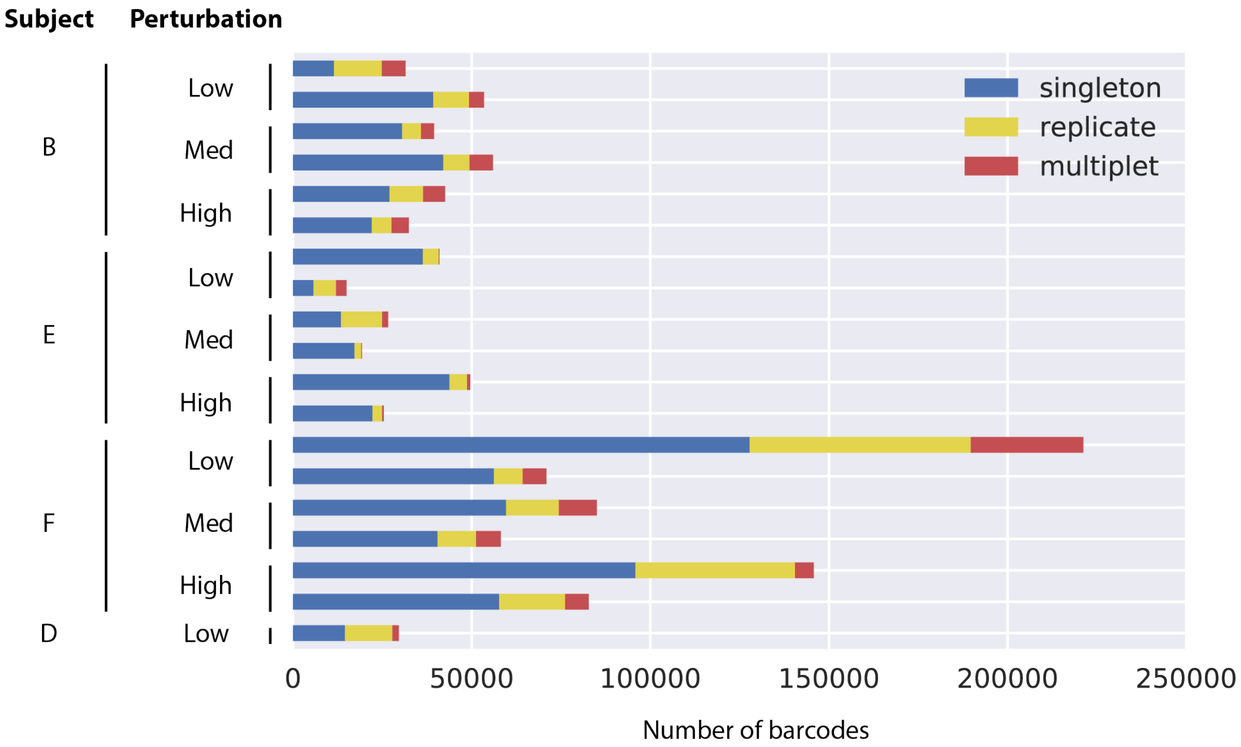

Figure B-1. Droplet barcoding showed variation between singleton, replicate, and multilpet OTU information. Subject ID and perturbation level are listed in columns on the left. Singletons represent instances of a single barcode mapping to a single read. Replicates include barcodes that match multiple reads, but all reads fall within the same OTU. Multiplets are unique barcodes that map to multiple OTUs, providing spatial co-localization information.

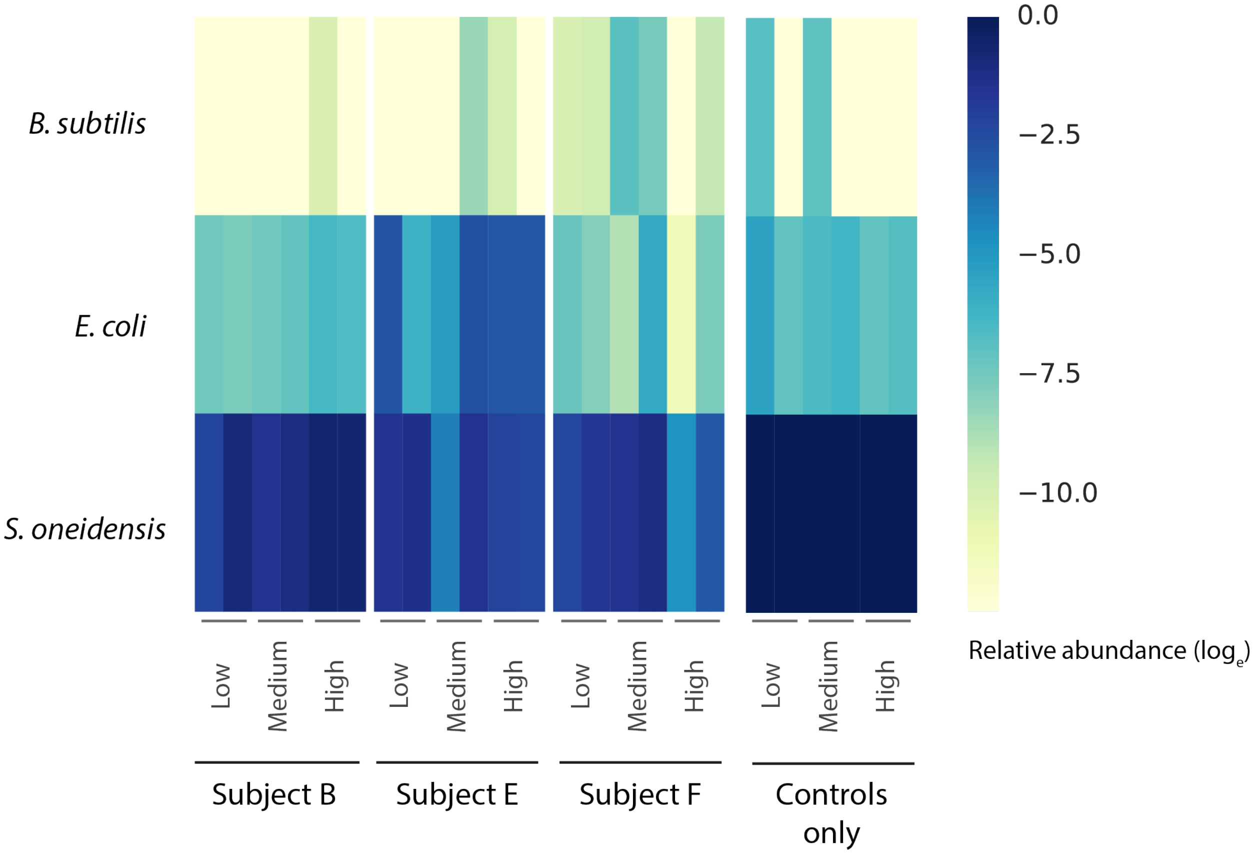

Figure B-2. Relative abundances of *B. subtilis*, *E. coli*, and *S. oneidensis* spike-ins within successfully amplified samples. Log-transformed relative abundance information was calculated using singleton barcode reads.

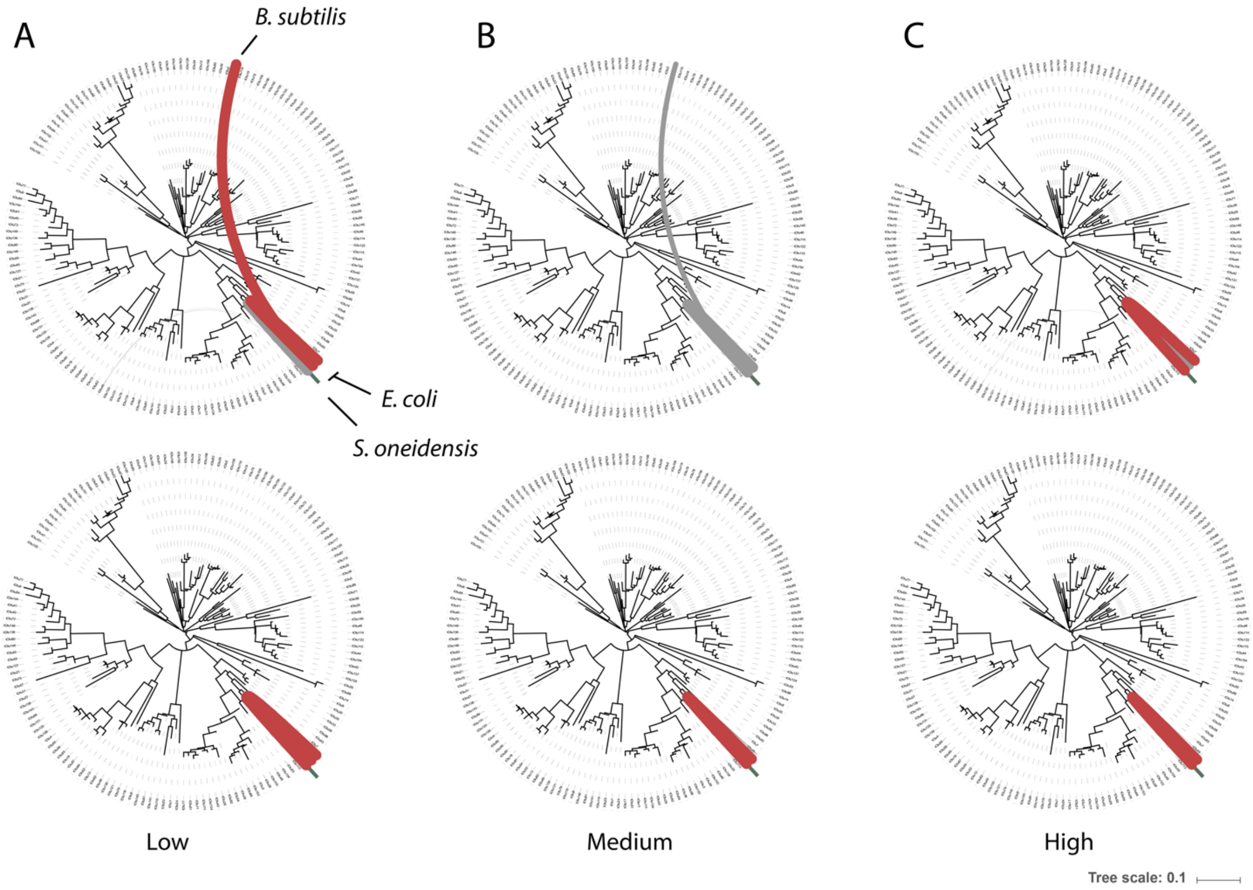

Figure B-3. Positive and negative control replicates at low, medium, and high levels of shear force. Columns A, B, and C represent low, medium, and high levels of shear force, respectively. The upper panel of column (A) demonstrates a single replicate that shows hypothesized behavior, with a significant connection between exclusively *E. coli* and *B. subtilis* spike-in positive controls. Remaining columns and replicates show consistent positive connections between the *S. oneidensis* negative control and the *E. coli* positive control.

B.3 Supplementary Tables

Table B-1. Sequencing primers used for clade-targeted and untargeting barcoding designs. Dashes indicate the join between two sequences in a bridge (in practice, there is no gap separating these primer components). Bold bases indicate Illumina overhangs.

Name	Description	Final concentration	Sequence (5’ -> 3’)
Streptococcus-targeted			
STb*	Fusion 1	1 µM	CGTTTGGAATTTCTCCGCTACCCA
TM7b	Fusion 2	1 µM	TKACCGCGGCTGCTG
STa*	Fusion bridge	10 nM	CTGAGCCAKRATCAAACTC-GCCTTTTGTAGAATGAACCGGCGA
STc*	Nested 1	300 nM	**ACACGACGCTCTTCCGATCTYRYR**TCACATGGTTTCGGGTCTA
TM7g	Nested 2	300 nM	**CGGCATTCCTGCTGAACCGCTCTTCCGATCT**GCGGCTGCTGGCACG
TM7i	Blocking 1	3.2 µM	TTTTTTTTTTGAGTTTGATYMTGGCTCAG/3SpC3/
TM7j	Blocking2	3.2 µM	TTTTTTTTTTCTGAGCCAKRATCAAACTC/3SpC3/
TM7-targeted			
TM7a	Fusion 1	1 µM	GAGTGACTGGGCGTAAA
TM7b	Fusion 2	1 µM	TKACCGCGGCTGCTG
TM7d	Fusion bridge	10 nM	CTGAGCCAKRATCAAACTC-CCCGTCAATTCCTTTATGTT
TM7f	Nested 1	300 nM	**ACACGACGCTCTTCCGATCTYRYR**GCGTAAAGAGTTGCGTAG
TM7g	Nested 2	300 nM	**CGGCATTCCTGCTGAACCGCTCTTCCGATCT**GCGGCTGCTGGCACG

TM7i	Blocking 1	3.2 μM	TTTTTTTTTTGAGTTTGATYMTGGCTCAG/3SpC3/
TM7j	Blocking 2	3.2 μM	TTTTTTTTTTCTGAGCCAKRATCAAACTC/3SpC3/
Untargeted barcoding			
fusion barcode	Droplet barcode	10 pM	CGGCACAATCTCGTCGCGTCGACACTCTTTCCCT**ACACGACGCTCTTCCGATCT**NNNNNNNNNNNNNNNNNNNNGATCATGACCCATTTGGAGAAGATG
barcode-Fw	Fusion 1	1 μM	CGGCACAATCTCGTCGCGTCG
1492R**	Fusion 2	1 μM	GGTTACCTTGTTACGACTT
barcodeR_519R**	Fusion bridge	10 nM	GWATTACCGCGGCKGCTG-CATCTTCTCCAAATGGGTCATGATC
illumina PCR for	Nested 1	300 nM	**ACACGACGCTCTTCCGATCT**
E786R**	Nested 2	300 nM	**CGGCATTCCTGCTGAACCGCTCTTCCGATCT**GGACTACHVGGGTWTCTAAT
U519F_block10**	Blocking 1	3.2 μM	TTTTTTTTTTGWATTACCGCGGCKGCTG/3SpC3/
U519R_block10**	Blocking 2	3.2 μM	TTTTTTTTTTCAGCMGCCGCGGTAATWC/3SpC3/

*Adapted from (76).

**Adapted from (56).

Table B-2. Taxonomic pairs which were linked to the same droplet barcode and filtered based on a Poisson model of random association based on relative abundance. B, E, F, D = different subjects. P = positive and negative control cells only. N = blank beads. Samples in red are depicted in Figure 3-5.

			Significant pairs including *S. oneidensis*	
Sample	Total pairs	Significant pairs ($p < 0.001$)	Significantly fewer connections than expected	Significantly more connections than expected
B, low	299	150	0	29
B, low	344	82	14	5
B, medium	249	53	7	3
B, medium	344	85	4	12
B, high	592	249	2	26
B, high	629	298	2	26
E, low	253	265	40	0
E, low	415	248	0	27
E, medium	145	29	2	0
E, medium	141	165	50	0
E, high	675	218	45	1
E, high	442	195	36	0
F, low	668	478	0	35
F, low	169	31	4	3
F, medium	298	84	4	15
F, medium	389	154	3	17
F, high	175	55	3	0
F, high	213	93	4	2
D, low	0	0	0	0
D, low	2	3	1	0
D, medium	0	2	2	0
D, medium	0	0	0	0
D, high	0	0	0	0
D, high	0	0	0	0
P, low	8	5	1	1
P, low	3	3	1	2
P, medium	2	1	1	0

P, medium	2	2	1	1
P, high	4	3	1	2
P, high	2	2	1	1
N, high	0	0	0	0
N, high	0	1	1	0

Table B-3. 16S rRNA gene amplicon library and final Illumina adapter addition primers. Universal 16S rRNA gene targeting segments are underlined.

Primer name	Primer sequence (5' -> 3')
PE-16S-V1V3-F	ACACGACGCTCTTCCGATCTYRYRGAGTTTGATYMTGGCTCAG
PE-16S-V1V3-R	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGCGGCTGCTGGCACG
PE-III-PCR-F-###	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE-IV-PCR-R-###	CAAGCAGAAGACGGCATACGAGATNNNNNNNNCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE-III-PCR-F_fusion	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
PE-IV-PCR-R_fusion	CAAGCAGAAGACGGCATACGAGATNNNNNNNNNCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Appendix C *Pseudomonas* genomics supplementary information

C.1 Supplementary Figures

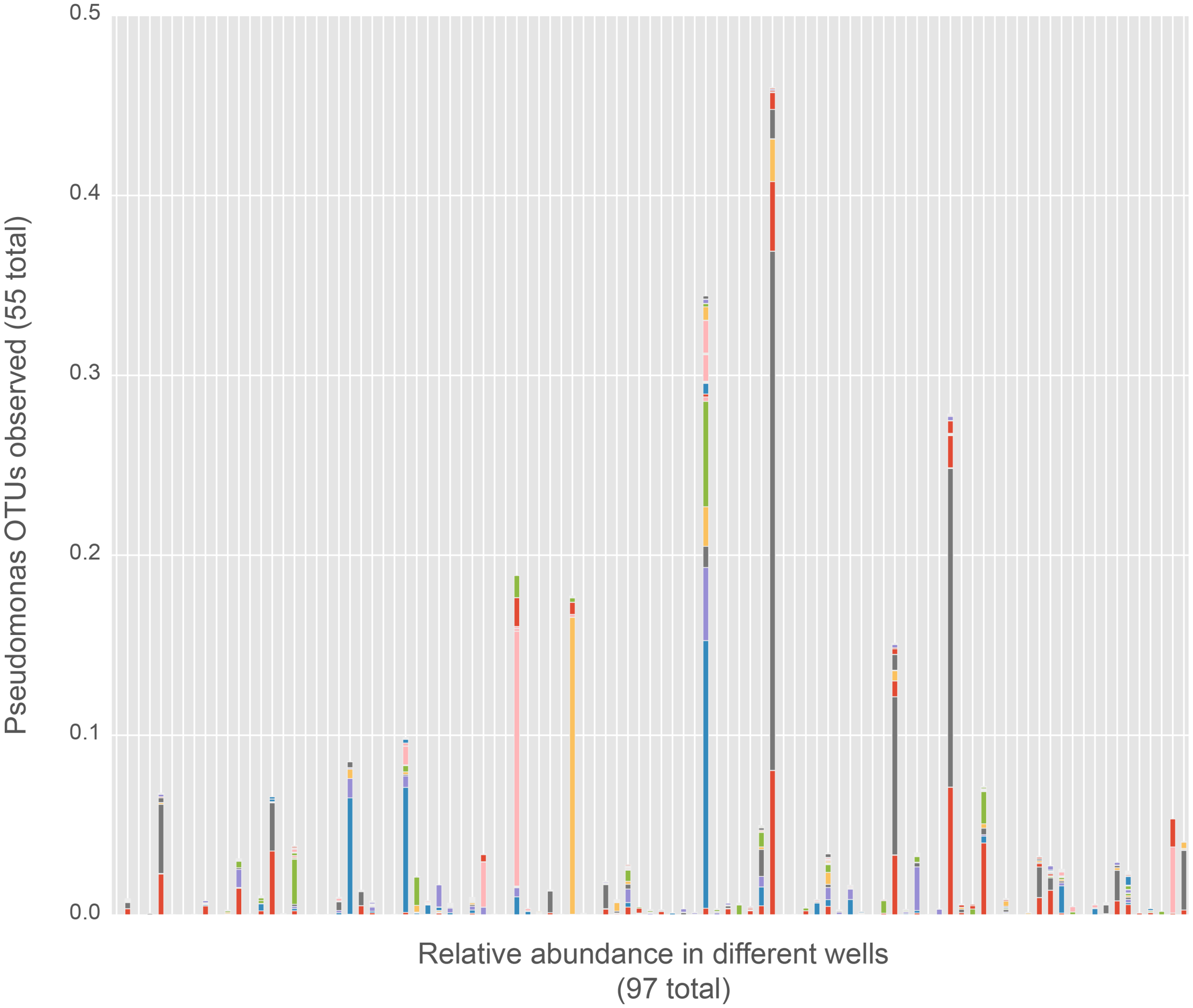

Figure C-1. *Pseudomonas* OTU presence in different samples from a 100-well survey **(96)**. Each color represents a unique OTU, and along the x axis are a total of 97 wells, with relative abundance summed from 221 groundwater samples at the Oak Ridge sampling site.

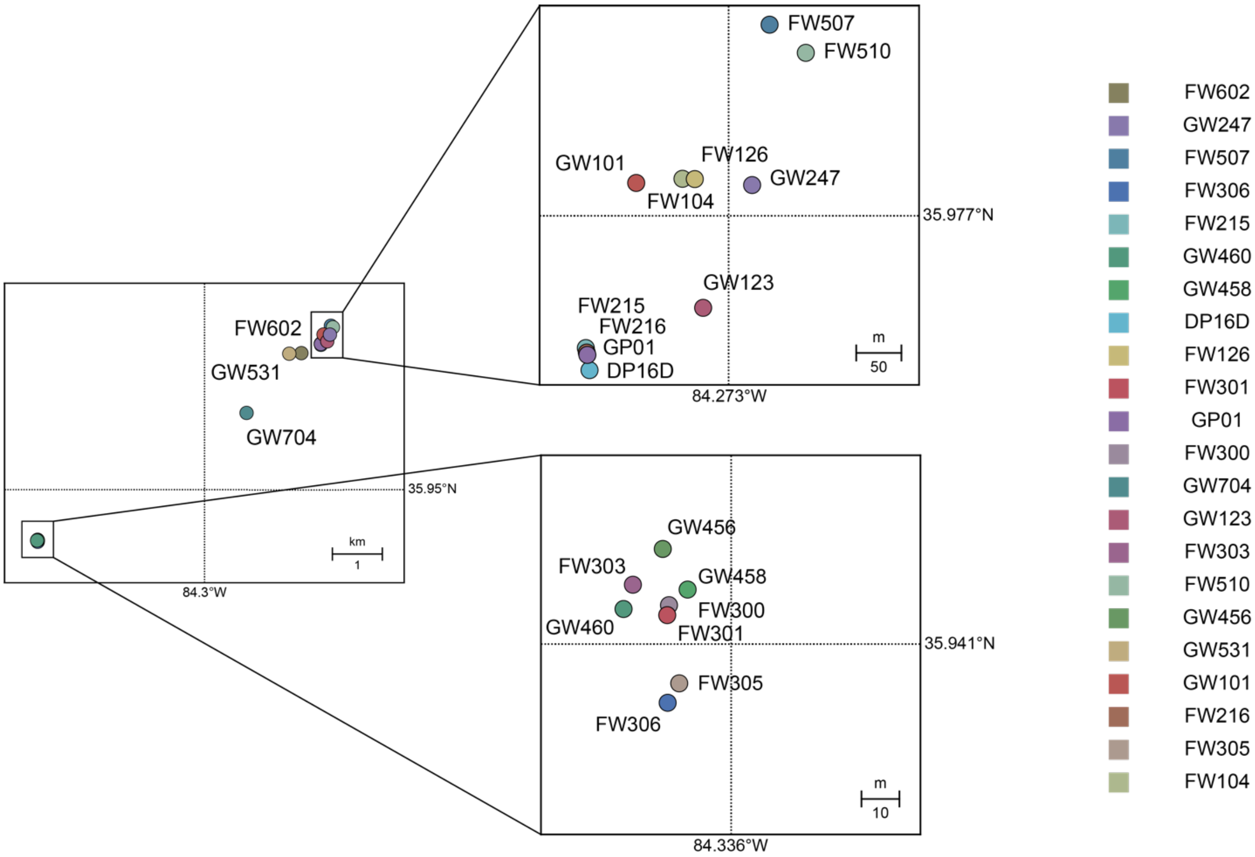

Figure C-2. GPS coordinates of the Oak Ridge, TN sampling sites. Each sampling well was assigned a unique color, and the direction of flow at the site is generally from Northeast to Southwest.

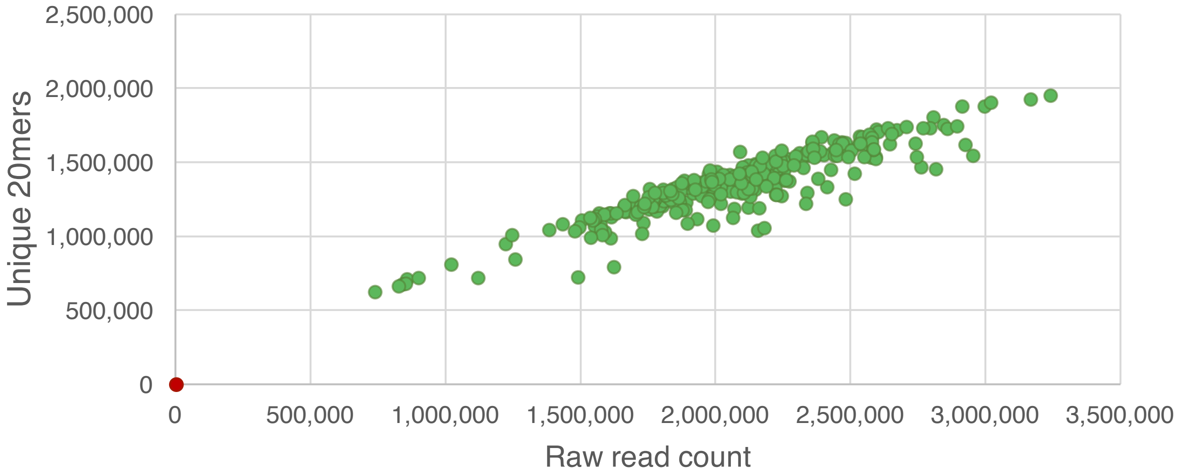

Figure C-3. Quality of low-volume whole-genome sequencing demonstrated with read counts and unique 20mer counts. Each dot represents a sequenced sample, and only one generated less than 500,000 reads (colored in red).

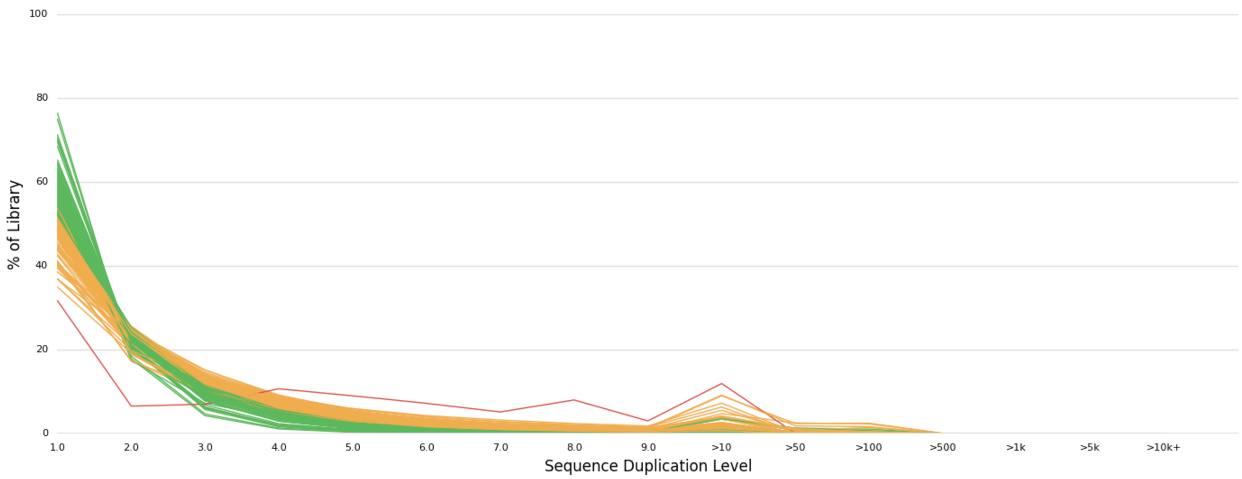

Figure C-4. Sequence duplicates generated through amplification of Nextera reactions. A duplication value of 1 indicates unique reads. Each line represents one genome, and colors represent duplication quality from high (green) to medium (yellow) and low (red). Summary generated by QUAST v4.5 and MultiQC v1.0 (146,147).

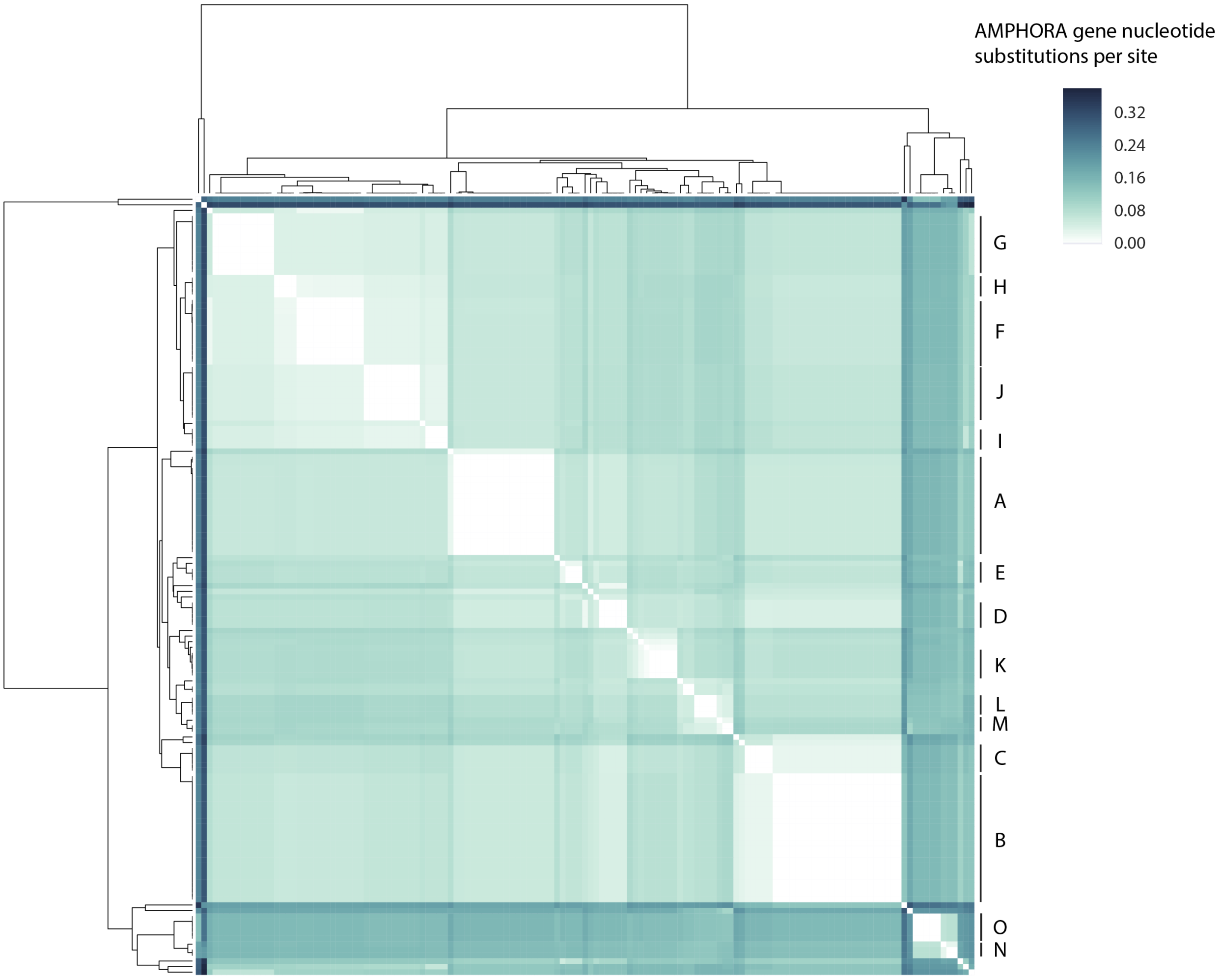

Figure C-5. Hierarchical clustering of nucleotide substitutions separating the *Pseudomonas* genome isolates. The color gradient displays nucleotide substitutions per site in a concatenated, masked alignment of AMPHORA genes identified with AMPHORA2. Hierarchical clustering was performed, and subgroups of strain-level genomes are labeled with letters corresponding to those depicted in Fig. 4-2A.

References

1. Yang X, Xie L, Li Y, and Wei C. More than 9,000,000 Unique Genes in Human Gut Bacterial Community: Estimating Gene Numbers Inside a Human Body. *PLoS One* **4**, e6074 (2009).

2. Locey KJ, and Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci USA* **113**, 5970–5975 (2016).

3. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

4. Guo C-J, Chang F-Y, Wyche TP, Backus KM, Acker TM, Funabashi M, *et al.* Discovery of Reactive Microbiota-Derived Metabolites that Inhibit Host Proteases. *Cell* **168**, 517–526 (2017).

5. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, *et al.* Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* **46**, 82–87 (2014).

6. Loh E, Salk JJ, and Loeb LA. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci USA* **107**, 1154–1159 (2010).

7. Chu ND, Clarke SA, Timberlake S, Polz MF, Grossman AD, and Alm EJ. A Mobile Element in mutS Drives Hypermutation in a Marine Vibrio. *MBio* **8**, e02045-16 (2017).

8. Polz MF, Alm EJ, and Hanage WP. Horizontal Gene Transfer and the Evolution of Bacterial and Archaeal Population Structure. *Trends Genet* **29**, 170–175 (2013).

9. Cahoon LA, and Seifert HS. Focusing homologous recombination: pilin antigenic variation in the pathogenic Neisseria. *Mol Microbiol* **81**, 1136–1143 (2011).

10. Preheim SP, Olesen SW, Spencer SJ, Materna A, Varadharajan C, Blackburn M, *et al.* Surveys, simulation and single-cell assays relate function and phylogeny in a lake ecosystem. *Nat Microbiol* **1**, 16130 (2016).

11. Morris JJ. Black Queen evolution: the role of leakiness in structuring microbial communities. *Trends Genet* **31**, 475–482 (2015).

12. Crowley DE, Wang YC, Reid CPP, and Szaniszlo PJ. Mechanisms of iron acquisition from siderophores by microorganisms and plants. *Plant Soil* **130**, 179–198 (1991).

13. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, *et al.* Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci USA* **112**, 244–249 (2015).

14. Pande S, Kaftan F, Lang S, Svatoš A, Germerodt S, and Kost C. Privatization of cooperative benefits stabilizes mutualistic cross-feeding interactions in spatially structured environments. *ISME J* **10**, 1413–1423 (2016).

15. von Ohle C, Gieseke A, Nistico L, Decker EM, DeBeer D, and Stoodley P. Real-time microsensor measurement of local metabolic activities in ex vivo dental biofilms exposed to sucrose and treated with chlorhexidine. *Appl Environ Microbiol* **76**, 2326–2334 (2010).

16. Ma S, and Banfield JF. Micron-scale Fe2+/Fe3+, intermediate sulfur species and O2 gradients across the biofilm–solution–sediment interface control biofilm organization. *Geochimica et Cosmochimica Acta* **75**, 3568–3580 (2011).

17. Snel B, Bork P, and Huynen MA. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**, 17–25 (2002).

18. Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, *et al.* Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**, 711–721 (2000).

19. Hinnebusch BJ, Rosso M-L, Schwan TG, and Carniel E. High-frequency conjugative transfer of antibiotic resistance genes to Yersinia pestis in the flea midgut. *Mol Microbiol* **46**, 349–354 (2002).

20. Xu L, Brito IL, Alm EJ, and Blainey PC. Virtual microfluidics for digital quantification and single-cell sequencing. *Nat Methods* **13**, 759–762 (2016).

21. Olsen I. Biofilm-specific antibiotic tolerance and resistance. *Eur J Clin Microbiol Infect Dis* **34**, 877–886 (2015).

22. Molin S, and Tolker-Nielsen T. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. *Curr Opin Biotechnol* **14**, 255–261 (2003).

23. Brileya KA, Camilleri LB, Zane GM, Wall JD, and Fields MW. Biofilm growth mode promotes maximum carrying capacity and community stability during product inhibition syntrophy. *Front Microbiol* **5**, 693 (2014).

24. Valm AM, Mark Welch JL, and Borisy GG. CLASI-FISH: principles of combinatorial labeling and spectral imaging. *Syst Appl Microbiol* **35**, 496–502 (2012).

25. Dechesne A, Pallud C, Debouzie D, Flandrois JP, Vogel TM, Gaudet JP, *et al.* A novel method for characterizing the microscale 3D spatial distribution of bacteria in soil. *Soil Biol and Biochem* **35**, 1537–1546 (2003).

26. Mark Welch JL, Rossetti BJ, Rieken CW, Dewhirst FE, and Borisy GG. Biogeography of a human oral microbiome at the micron scale. *Proc Natl Acad Sci USA* **113**, E791-800 (2016).

27. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, *et al.* Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. *Nat Genet* **45**, 1176–1182 (2013).

28. Stepanauskas R, and Sieracki ME. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* **104**, 9052–9057 (2007).

29. Siegl A, and Hentschel U. PKS and NRPS gene clusters from microbial symbiont cells of marine sponges by whole genome amplification. *Environ Microbiol Rep* **2**, 507–513 (2010).

30. Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME, *et al.* High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**, 113–123 (2012).

31. Bayer K, Scheuermayer M, Fieseler L, and Hentschel U. Genomic mining for novel $FADH_2$-dependent halogenases in marine sponge-associated microbial consortia. *Mar Biotechnol* **15**, 63–72 (2013).

32. Ottesen EA, Hong JW, Quake SR, and Leadbetter JR. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* **314**, 1464–1467 (2006).

33. Zeng Y, Novak R, Shuga J, Smith MT, and Mathies RA. High-performance single cell genetic analysis using microfluidic emulsion generator arrays. *Anal Chem* **82**, 3183–3190 (2010).

34. Tadmor AD, Ottesen EA, Leadbetter JR, and Phillips R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**, 58–62 (2011).

35. Gieseke A, Bjerrum L, Wagner M, and Amann R. Structure and activity of multiple nitrifying bacterial populations co-existing in a biofilm. *Environ Microbiol* **5**, 355–369 (2003).

36. Valm AM, Mark Welch JL, Rieken CW, Hasegawa Y, Sogin ML, Oldenbourg R, *et al.* Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proc Natl Acad Sci USA* **108**, 4152–4157 (2011).

37. Baptista JDC, Lunn M, Davenport RJ, Swan DL, Read LF, Brown MR, *et al.* Agreement between amoA gene-specific quantitative PCR and fluorescence in situ hybridization in the measurement of ammonia-oxidizing bacteria in activated sludge. *Appl Environ Microbiol* **80**, 5901–5910 (2014).

38. Turner DJ, and Hurles ME. High-throughput haplotype determination over long distances by haplotype fusion PCR and ligation haplotyping. *Nat Protocols* **4**, 1771–1783 (2009).

39. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* **43**, 2507–2515 (2013).

40. Yon J, and Fried M. Precise gene fusion by PCR. *Nucleic Acids Res* **17**, 4895 (1989).

41. Tamminen MV, and Virta MPJ. Single gene-based distinction of individual microbial genomes from a mixed population of microbial cells. *Front Microbiol* **6**, 195 (2015).

42. Holmes DL, and Stellwagen NC. Estimation of polyacrylamide gel pore size from Ferguson plots of linear DNA fragments. II. Comparison of gels with different crosslinker concentrations, added agarose and added linear polyacrylamide. *Electrophoresis* **12**, 612–619 (1991).

43. Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong S-H, *et al.* The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol Cell* **44**, 252–264 (2011).

44. Leloup J, Quillet L, Oger C, Boust D, and Petit F. Molecular quantification of sulfate-reducing microorganisms (carrying dsrAB genes) by competitive PCR in estuarine sediments. *FEMS Microbiol Ecol* **47**, 207–214 (2004).

45. Wetmur JG, Kumar M, Zhang L, Palomeque C, Wallenstein S, and Chen J. Molecular haplotyping by linking emulsion PCR: analysis of paraoxonase 1 haplotypes and phenotypes. *Nucl Acids Res* **33**, 2615–2619 (2005).

46. Müller AL, Kjeldsen KU, Rattei T, Pester M, and Loy A. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *ISME J* **9**, 1152–1165 (2015).

47. Price MN, Dehal PS, and Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

48. Muyzer G, and Stams AJM. The ecology and biotechnology of sulphate-reducing bacteria. *Nat Rev Micro* **6**, 441–454 (2008).

49. Watanabe T, Kojima H, and Fukui M. Draft Genome Sequence of a Psychrotolerant Sulfur-Oxidizing Bacterium, Sulfuricella denitrificans skB26, and Proteomic Insights into Cold Adaptation. *Appl Environ Microbiol* **78**, 6545–6549 (2012).

50. Kallmeyer J, Smith DC, Spivack AJ, and D'Hondt S. New cell extraction procedure applied to deep subsurface sediments. *Limnol Oceanogr Methods* **6**, 236–245 (2008).

51. Liu J, Li J, Feng L, Cao H, and Cui Z. An improved method for extracting bacteria from soil for high molecular weight DNA recovery and BAC library construction. *J Microbiol* **48**, 728–733 (2011).

52. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, and Griffiths AD. Amplification of complex gene libraries by emulsion PCR. *Nat Meth* **3**, 545–550 (2006).

53. Wagner M, Loy A, Klein M, Lee N, Ramsing NB, Stahl DA, *et al.* Functional marker genes for identification of sulfate-reducing prokaryotes. *Meth Enzymol* **397**, 469–489 (2005).

54. Wagner M, Roger AJ, Flax JL, Brusseau GA, and Stahl DA. Phylogeny of dissimilatory sulfite reductases supports an early origin of sulfate respiration. *J Bacteriol* **180**, 2975–2982 (1998).

55. Giloteaux L, Goñi-Urriza M, and Duran R. Nested PCR and new primers for analysis of sulfate-reducing bacteria in low-cell-biomass environments. *Appl Environ Microbiol* **76**, 2856–2865 (2010).

56. Lane DJ. 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systemantics*, pp 115–175 (Wiley & Sons, 1991).

57. Cradic KW, Wells JE, Allen L, Kruckeberg KE, Singh RJ, and Grebe SKG. Substitution of 3'-phosphate cap with a carbon-based blocker reduces the possibility of fluorescence resonance energy transfer probe failure in real-time PCR assays. *Clin Chem* **50**, 1080–1082 (2004).

58. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* **7**, 335–336 (2010).

59. Baker GC, Smith JJ, and Cowan DA. Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods* **55**, 541–555 (2003).

60. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069–5072 (2006).

61. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl Acids Res* **41**, D590–D596 (2013).

62. Pruesse E, Peplies J, and Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).

63. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

64. Rice P, Longden I, and Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277 (2000).

65. Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

66. Letunic I, and Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucl Acids Res* **39**, W475–W478 (2011).

67. Donlan RM, and Costerton JW. Biofilms: Survival Mechanisms of Clinically Relevant Microorganisms. *Clin Microbiol Rev* **15**, 167–193 (2002).

68. Periasamy S, and Kolenbrander PE. Mutualistic biofilm communities develop with Porphyromonas gingivalis and initial, early, and late colonizers of enamel. *J Bacteriol* **191**, 6804–6811 (2009).

69. Edlund A, Yang Y, Yooseph S, Hall AP, Nguyen DD, Dorrestein PC, *et al.* Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. *ISME J* **9**, 2605–2619 (2015).

70. Farnelid HM, Turk-Kubo KA, and Zehr JP. Identification of Associations between Bacterioplankton and Photosynthetic Picoeukaryotes in Coastal Waters. *Front Microbiol* **7**, 339 (2016).

71. Hartmann M, Zubkov MV, Scanlan DJ, and Lepère C. In situ interactions between photosynthetic picoeukaryotes and bacterioplankton in the Atlantic Ocean: evidence for mixotrophy. *Environ Microbiol Rep* **5**, 835–840 (2013).

72. Spencer SJ, Tamminen MV, Preheim SP, Guo MT, Briggs AW, Brito IL, *et al.* Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J* **10**, 427–436 (2016).

73. Pregibon DC, and Doyle PS. Optimization of encoded hydrogel particles for nucleic acid quantification. *Anal Chem* **81**, 4873–4881 (2009).

74. Kolenbrander PE, Palmer RJ, Periasamy S, and Jakubovics NS. Oral multispecies biofilm development and the key role of cell–cell distance. *Nat Rev Micro* **8**, 471–480 (2010).

75. Ruhl S, Eidt A, Melzl H, Reischl U, and Cisar JO. Probing of Microbial Biofilm Communities for Coadhesion Partners. *Appl Environ Microbiol* **80**, 6583–6590 (2014).

76. Moore MS, McCarroll MG, McCann CD, May L, Younes N, and Jordan JA. Direct Screening of Blood by PCR and Pyrosequencing for a 16S rRNA Gene Target from Emergency Department and Intensive Care Unit Patients Being Evaluated for Bloodstream Infection. *J Clin Microbiol* **54**, 99–105 (2016).

77. Cassels FJ, and London J. Isolation of a coaggregation-inhibiting cell wall polysaccharide from Streptococcus sanguis H1. *J Bacteriol* **171**, 4019–4025 (1989).

78. Chalmers NI, Palmer RJ, Cisar JO, and Kolenbrander PE. Characterization of a Streptococcus sp.-Veillonella sp. Community Micromanipulated from Dental Plaque. *J Bacteriol* **190**, 8145–8154 (2008).

79. Palmer RJ, Shah N, Valm A, Paster B, Dewhirst F, Inui T, *et al.* Interbacterial Adhesion Networks within Early Oral Biofilms of Single Human Hosts. *Appl Environ Microbiol* **83**, e00407-17 (2017).

80. Maeda K, Nagata H, Kuboniwa M, Ojima M, Osaki T, Minamino N, *et al.* Identification and Characterization of Porphyromonas gingivalis Client Proteins That Bind to Streptococcus oralis Glyceraldehyde-3-Phosphate Dehydrogenase. *Infect Immun* **81**, 753–763 (2013).

81. Schulze-Schweifing K, Banerjee A, and Wade WG. Comparison of bacterial culture and 16S rRNA community profiling by clonal analysis and pyrosequencing for the characterization of the dentine caries-associated microbiome. *Front Cell Infect Microbiol* **4**, 164 (2014).

82. Kolenbrander PE, Andersen RN, and Holdeman LV. Coaggregation of oral Bacteroides species with other bacteria: central role in coaggregation bridges and competitions. *Infect Immun* **48**, 741–746 (1985).

83. Soro V, Dutton LC, Sprague SV, Nobbs AH, Ireland AJ, Sandy JR, *et al.* Axenic Culture of a Candidate Division TM7 Bacterium from the Human Oral Cavity and Biofilm Interactions with Other Oral Bacteria. *Appl Environ Microbiol* **80**, 6480–6489 (2014).

84. Lens P, O'Flaherty V, Moran AP, Stoodley P, and Mahony T. *Biofilms in Medicine, Industry and Environmental Biotechnology* (IWA Publishing, 2003).

85. Bodenmiller D, Toh E, and Brun YV. Development of surface adhesion in Caulobacter crescentus. *J Bacteriol* **186**, 1438–1447 (2004).

86. Levi A, and Jenal U. Holdfast formation in motile swarmer cells optimizes surface attachment during Caulobacter crescentus development. *J Bacteriol* **188**, 5315–5318 (2006).

87. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protocols* **10**, 442–458 (2015).

88. Turchaninova MA, Britanova OV, Bolotin DA, Shugay M, Putintseva EV, Staroverov DB, *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur J Immunol* **43**, 2507–2515 (2013).

89. Zhang J, Kobert K, Flouri T, and Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

90. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, and Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)* **2010**, baq013 (2010).

91. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D, *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun* **4**, 2120 (2013).

92. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

93. Wang D, Boukhalfa H, Ware DS, and Daligault HE. Draft Genome Sequence of a Chromium-Reducing Strain, Pseudomonas fluorescens S613, Isolated from a Chromium-Contaminated Aquifer in Los Alamos, New Mexico. *Genome Announc* **5**, e00241-17 (2017).

94. Nielsen TK, Kot W, Sørensen SR, and Hansen LH. Draft Genome Sequence of MCPA-Degrading Sphingomonas sp. Strain ERG5, Isolated from a Groundwater Aquifer in Denmark. *Genome Announc* **3**, e01529-14 (2015).

95. Wu X, Holmfeldt K, Hubalek V, Lundin D, Åström M, Bertilsson S, *et al.* Microbial metagenomes from three aquifers in the Fennoscandian shield terrestrial deep biosphere reveal metabolic partitioning among populations. *ISME J* **10**, 1192–1203 (2016).

96. Smith MB, Rocha AM, Smillie CS, Olesen SW, Paradis C, Wu L, *et al.* Natural bacterial communities serve as quantitative geochemical biosensors. *MBio* **6**, e00326-315 (2015).

97. Besmer MD, Epting J, Page RM, Sigrist JA, Huggenberger P, and Hammes F. Online flow cytometry reveals microbial dynamics influenced by concurrent natural and operational events in groundwater used for drinking water treatment. *Sci Rep* **6**, 38462 (2016).

98. Whitman WB, Coleman DC, and Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**, 6578–6583 (1998).

99. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**, 13219 (2016).

100. Keller AH, Schleinitz KM, Starke R, Bertilsson S, Vogt C, and Kleinsteuber S. Metagenome-Based Metabolic Reconstruction Reveals the Ecophysiological Function of Epsilonproteobacteria in a Hydrocarbon-Contaminated Sulfidic Aquifer. *Front Microbiol* **6**, 1396 (2015).

101. Chakraborty R, Woo H, Dehal P, Walker R, Zemla M, Auer M, *et al.* Complete genome sequence of Pseudomonas stutzeri strain RCH2 isolated from a Hexavalent Chromium [Cr(VI)] contaminated site. *Stand Genomic Sci* **12**, 23 (2017).

102. McRobb E, Sarovich DS, Price EP, Kaestli M, Mayo M, Keim P, *et al.* Tracing melioidosis back to the source: using whole-genome sequencing to investigate an outbreak originating from a contaminated domestic water supply. *J Clin Microbiol* **53**, 1144–1148 (2015).

103. Russell JA, León-Zayas R, Wrighton K, and Biddle JF. Deep Subsurface Life from North Pond: Enrichment, Isolation, Characterization and Genomes of Heterotrophic Bacteria. *Front Microbiol* **7**, 678 (2016).

104. Wilkins MJ, Kennedy DW, Castelle CJ, Field EK, Stepanauskas R, Fredrickson JK, *et al.* Single-cell genomics reveals metabolic strategies for microbial growth and survival in an oligotrophic aquifer. *Microbiology* **160**, 362–372 (2014).

105. Kumar HKS, Gan HM, Tan MH, Eng WWH, Barton HA, Hudson AO, *et al.* Genomic characterization of eight Ensifer strains isolated from pristine caves and a whole genome phylogeny of Ensifer (Sinorhizobium). *J Genomics* **5**, 12–15 (2017).

106. Preheim SP, Perrotta AR, Martin-Platero AM, Gupta A, and Alm EJ. Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**, 6593–6603 (2013).

107. Vaccaro BJ, Lancaster WA, Thorgersen MP, Zane GM, Younkin AD, Kazakov AE, *et al.* Novel Metal Cation Resistance Systems from Mutant Fitness Analysis of Denitrifying Pseudomonas stutzeri. *Appl Environ Microbiol* **82**, 6046–6056 (2016).

108. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, and Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–55 (2015).

109. Wu M, and Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

110. Zha D, Xu L, Zhang H, and Yan Y. The two-component GacS-GacA system activates lipA translation by RsmE but not RsmA in Pseudomonas protegens Pf-5. *Appl Environ Microbiol* **80**, 6627–6637 (2014).

111. Liao CH, McCallus DE, Fett WF, and Kang Y. Identification of gene loci controlling pectate lyase production and soft-rot pathogenicity in Pseudomonas marginalis. *Can J Microbiol* **43**, 425–431 (1997).

112. Yu X, Chen M, Jiang Z, Hu Y, and Xie Z. The Two-Component Regulators GacS and GacA Positively Regulate a Nonfluorescent Siderophore through the Gac/Rsm Signaling Cascade in High-Siderophore-Yielding Pseudomonas sp. Strain HYS. *J Bacteriol* **196**, 3259–3270 (2014).

113. Mukhopadhyay S, Audia JP, Roy RN, and Schellhorn HE. Transcriptional induction of the conserved alternative sigma factor RpoS in Escherichia coli is dependent on BarA, a probable two-component regulator. *Mol Microbiol* **37**, 371–381 (2000).

114. Lange R, Fischer D, and Hengge-Aronis R. Identification of transcriptional start sites and the role of ppGpp in the expression of rpoS, the structural gene for the sigma S subunit of RNA polymerase in Escherichia coli. *J Bacteriol* **177**, 4676–4680 (1995).

115. Kojic M, and Venturi V. Regulation of rpoS Gene Expression in Pseudomonas: Involvement of a TetR Family Regulator. *J Bacteriol* **183**, 3712–3720 (2001).

116. Venturi V. Control of rpoS transcription in Escherichia coli and Pseudomonas: why so different? *Mol Microbiol* **49**, 1–9 (2003).

117. Koehorst JJ, van Dam JCJ, van Heck RGA, Saccenti E, dos Santos VAPM, Suarez-Diez M, *et al.* Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data. *Sci Rep* **6**, 38699 (2016).

118. Pernestig AK, Melefors O, and Georgellis D. Identification of UvrY as the cognate response regulator for the BarA sensor kinase in Escherichia coli. *J Biol Chem* **276**, 225–231 (2001).

119. Parkins MD, Ceri H, and Storey DG. Pseudomonas aeruginosa GacA, a factor in multihost virulence, is also essential for biofilm formation. *Mol Microbiol* **40**, 1215–1226 (2001).

120. Coates JD, Lonergan DJ, Philips EJP, Jenter H, and Lovley DR. Desulfuromonas palmitatis sp. nov., a marine dissimilatory Fe(III) reducer that can oxidize long-chain fatty acids. *Arch Microbiol* **164**, 406–413 (1995).

121. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

122. Bolger AM, Lohse M, and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

123. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455–477 (2012).

124. Edgar R. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv* 74161 (2016).

125. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, and Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108 (2007).

126. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).

127. Talavera G, and Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577 (2007).

128. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

129. Treangen TJ, Ondov BD, Koren S, and Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* **15**, 524 (2014).

130. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

131. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997v2 (2013).

132. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

133. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

134. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, *et al.* Integrative Genomics Viewer. *Nat Biotechnol* **29**, 24–26 (2011).

135. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, and Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).

136. Rhoads A, and Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

137. Pamp SJ, Harrington ED, Quake SR, Relman DA, and Blainey PC. Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome Res* **22**, 1107–1119 (2012).

138. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, *et al.* Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc* **12**, 44–73 (2017).

139. Lan F, Demaree B, Ahmed N, and Abate AR. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat Biotechnol* **35**, 640–646 (2017).

140. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302–308 (2017).

141. Vakatov D. *NCBI C toolkit* (National Center for Biotechnology Information, U.S. National Library of Medicine, 2013).

142. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, *et al.* Unlocking short read sequencing for metagenomics. *PLoS One* **5**, e11840 (2010).

143. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537–7541 (2009).

144. Smith JL, Goldberg JM, and Grossman AD. Complete Genome Sequences of Bacillus subtilis subsp. subtilis Laboratory Strains JH642 (AG174) and AG1839. *Genome Announc* **2**, e00663-14 (2014).

145. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 81257 (2016).

146. Gurevich A, Saveliev V, Vyahhi N, and Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

147. Ewels P, Magnusson M, Lundin S, and Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).