DISSERTATION

METAGENOMIC INSIGHTS INTO MICROBIAL COLONIZATION & PERSISTENCE IN SUBSURFACE FRACTURED SHALES

Submitted by

Kaela K. Amundson

Department of Soil & Crop Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor Michael J. Wilkins

Kelly C. Wrighton
Thomas Borch
Matthew Ross

ABSTRACT

METAGENOMIC INSIGHTS INTO MICROBIAL COLONIZATION & PERSISTENCE IN SUBSURFACE FRACTURED SHALES

Microorganisms are pervasive yet important components of hydraulically fractured shale systems. Subsurface shales harbor oil & gas and require unconventional techniques, such as hydraulic fracturing, to access these trapped hydrocarbons. Shale microbiomes are of crucial importance as they can directly impact the recovery of oil & gas and associated infrastructure. The overarching theme of this dissertation was to characterize the metabolisms and key traits that underpin the colonization and persistence of fractured shale microbiomes using a multi-omic approach to better understand the microbial impact on this important ecosystem.

In Chapter 1, I first discuss the importance of subsurface shales as an important energy reserve, summarize what is known about microorganisms in these ecosystems, and highlight the strength of using a metagenomic approach to studying shale microbiomes. Subsurface shales are heterogeneous – varying in their mineral content, temperature, and other physiochemical conditions. The microbial communities that persist can have substantial impact on the fractured shale ecosystem and contribute to common challenges in hydrocarbon recovery such as corrosion, souring, and bioclogging. The literature review presented here highlights the need to study the functional potential of shale microbiomes as most studies have mostly focused solely on taxonomic composition of persisting microbial communities, and a vast majority of these studies have focused on samples from wells in the Appalachian Basin. However, functional potential of shale microbiomes across a variety of physiochemical conditions must be considered

in order to gain an understanding of the role of microorganisms and what possible influences they may have on hydraulically fractured shales systems. Here, I highlight the need (1) to study the whole community at a functional scale and (2) apply a metagenomic approach to a variety of less characterized shale basins to gain a holistic understanding of shale microbiomes and the effects they may have on the broader ecosystem.

In Chapter 2, I apply this metagenomic approach to study the persisting shale microbiomes of three fractured shale wells in the Anadarko Basin – a western shale basin characterized by elevated temperature and salinity. No studies using metagenomics had been applied to shale basins in the western United States prior to this research. We sampled five wells in the Anadarko basin over a timeseries >500 days and preformed NMR metabolomics and metagenomic sequencing to uncover the dominant metabolisms, community composition, and other functional traits of the Anadarko shale microbiome. This system was dominated by a fermentative microbial community and a less-abundant sub-community of inferred sulfate reducing microorganisms. Using paired NMR metabolomics and metagenomics, I demonstrated how many fermentative microorganisms have the potential to degrade common complex polymers, such as guar gum, and have potential to produce organic acids that may serve as electron donors for sulfate and thiosulfate reducing microorganisms. Thus, in this study I provided a framework for how carbon may move through the closed fractured shale ecosystem to sustain the microbial community. Finally, I investigated viral presence and diversity across all thirty-six metagenomes and found that inputs were large sources of viral diversity, but that only an extremely small proportion of viruses recovered from produced fluids were genetically similar to viruses previously reported from fractured shales. I observed that a majority of the dominant and persisting genomes encoded a CRISPR-Cas viral defense system, likely in response to the

viral community. This highlights viral defense as another key trait for persisting microorganisms, as viruses are the only predators to bacteria and archaea in fractured shale ecosystems. Overall, this study expanded our knowledge of sulfate and thiosulfate reducing microorganisms in fractured shales, demonstrated the potential for common chemical inputs such as guar gum to be utilized by shale microbiomes, and highlighted how other key traits, such as CRISPR-Cas viral defense systems, may be a crucial trait for persisting shale microbiomes.

Building on results from viral analyses in Chapter 2, in Chapter 3 I next sought to investigate the temporal dynamics between hosts and viruses to better understand the role of microbial defense against viruses in fractured shale ecosystems. To do this, I sampled six shale wells in the Denver-Julesburg Basin for >800 days, performed metagenomic sequencing, and identified host (bacterial & archaeal) and viral genomes from this data. I observed evidence of ongoing host defense to viral predation at both the community and genome-level through quantifying spacers from CRISPR arrays in metagenomic reads and MAGs. Through these analyses leveraging timeseries sampling and age differences between the shale wells, I provided evidence that suggested migration toward CRISPR arrays that may be more efficient at protecting the microbial host against a wider suite of viruses. Finally, I observed a temporal increase in host-viral co-existence in the closed, fractured shale ecosystem – suggesting the CRISPR defense does not entirely protect against viral predation.

Chapter 4 ultimately leverages the approaches, insights, and data gained from Chapters 2 and 3 to study shale microbiomes at a cross-basin, geographic scale. Here, I collected samples from many collaborators who have previously worked in shale systems, performed metagenomic sequencing, and processed all samples in a standardized pipeline to build a comprehensive genomic shale database. In total, this database contains 978 unique MAGs and >7 million unique

genes recovered from 209 metagenomic samples obtained from 36 fractured shale wells spanning eleven shale basins from North America, China, and the United Kingdom. In this chapter I analyze the functional potential of shale microbiomes at a genome-resolved level to better understand the geographic distribution of microbial metabolisms and other key traits that likely contribute to colonization and persistence of microorganisms. Here I also leveraged bioinformatic tools to build a custom annotation summary toolkit to process and analyze the large amount of sequencing data for traits of interest. The complete absence of a taxonomic core microbiome across shale basins illustrated in this chapter underscores the necessity of a genome-resolved and functional approach to studying shale microbiomes. Results from analyzing shale microbiomes at this scale could ultimately help to inform microbial management of fractured shale systems.

The final chapter of this dissertation (Chapter 5) summarizes the key findings of my research into fractured shale microbiomes, and the mechanisms that may promote microbial colonization, persistence, and survival in these relatively harsh and economically relevant ecosystems. In this chapter I conclude this work by discussing future directions and lingering knowledge gaps for studying fractured shale microbiomes, as well as implications of these findings for other subsurface engineered ecosystems. Ultimately, this body of work contributes a substantial amount of new and informative insights into the functional potentials of persisting shale microbiomes across broad geographic scales.

ACKNOWLEDGMENTS

The past five years have taught me the power of becoming comfortable being uncomfortable. Whether it was presenting to an intimidating audience or traveling internationally for the first time, the journey to earning this degree has given me so many amazing, unique, and uncomfortable experiences that I am forever grateful for. The culmination of work presented here is a direct result of the opportunities that were provided by supportive mentors, the scientific, professional, and personal growth that came from feeling uncomfortable, and the support from family, friends, mentors, and peers as I navigated those experiences.

First, a profound thank you to my advisor, **Dr. Mike Wilkins**. Without doubt I am here, five years later, completing this degree because of your guidance, encouragement, and perhaps most of all – patience – as I grew into the scientist I am today. I am so grateful for the various side projects and leadership opportunities that you encouraged and provided for me, and the fun science adventures I've gotten to partake in because of that. Most of all, thank you for being an example of what it is to be a kind, centered, and unfailingly supportive mentor. Whether I was excited to share new results or discouraged and spinning in circles, I always left our meetings feeling grounded and knowing which direction to move forward. I believe this to be one of the best lessons you've taught me in science and mentorship, and I am grateful to have had the opportunity to learn from you (and with you) over the past five years.

I'd also like to thank my committee for their contributions to this work. To **Dr. Kelly Wrighton** – thank you for pushing me, providing opportunities I would have otherwise not had, and for your support and guidance along the way. I have grown because of your contributions to this dissertation research, and I am grateful to know you as a mentor and collaborator.

Additionally, thank you to **Dr. Thomas Borch** for your expertise in hydraulically fractured shale systems, and **Dr. Matthew Ross** for your expertise in ecology and genuine kindness and support.

Over the past five years I have been lucky to be surrounded by amazing labmates in the Wilkins and Wrighton groups who've also been mentors, peers, and friends. To my first grad school mentor, **Dr. Mikayla Borton**, thank you for teaching me about the shale project, how to use and trouble shoot computational tools to process metagenomic data on the command line, and answering my millions of questions in the process. You helped teach me the fundamentals crucial for this research, and for that I am grateful. I'd also like to thank **Reb Daly** for training me in the wet lab, on shale, and for providing a primer on viruses and CRISPR arrays. An immense thank you to **Dr. Bridget McGivern** for your thoughtful, measured advice and always being available to talk through problems whether it was related to science or life. From starting as roommates to growing as friends – it has been a journey and I have appreciated your friendship through it all. The past five years would not have been the same without **Amelia Nelson**, who joined me in the Wilkins lab just eight months after I started. Thank you for all the bobs happy hours and venting sessions, especially through years of outside only seating and work from home due to COVID. Finally, to **Dr. Josue Rodriguez-Ramos**, **Dr. Emily Bechtold**, **Julie Fowler**, and **Raegan Paul**, thank you for being such kind, supportive, and fun lab mates. I am lucky to know you all as both friends and colleagues.

I would not have started this journey without the inspiration and support of many undergraduate mentors. To **Dr. Andrew Bent** and the many graduate students that I met during my four years in the Bent lab, mostly **Dr. Adam Bayless**, **Dr. Katelyn Butler**, and **Dr. Ryan Zapotocny**, as well as **Dr. Trina McMahon** and former lab members **Dr. Alex Linz** and **Dr. Elizabeth McDaniel –** thank you thank you for inspiring me to pursue a PhD, for answering my

many, many, questions about the graduate school application process, and being excellent role models and mentors in science and life.

There are so many more friendships– beginning before graduate school and during – that have supported me through the various stages of completing this work, and it would take far too much space and time to list them all. To my oldest friend, **Nicole Schulz**, thank you for being a steadfast friend who I know will always be there despite the miles, time apart, and many life changes. To **Abby Waller**, thank you for being the best listener and for being an inspiring example of strength and perseverance. To **Konnor Heyde**, thank you for your sincere friendship, for always being one of my biggest cheerleaders, and for picking up every tearful late night phone call. I'd also like to thank the **SOCR grads** for the community we built and fostered together, even through years of COVID. Finally, the past few years would not have been the same without the happy hours, trivia nights, and other adventures with **Gina Cerimele**, **Tony Schuh**, **Kristen Johnson**, **Brad Pakish**, **Allison Belfield, and so many others**.

Of course, there are not enough words to express my appreciation to my parents, **John & Laura Amundson**. You raised three strong daughters who know the power of hard work because you set an example of this yourselves. You gave me my stubbornness for high standards and taught me that the hardest things are always worth the effort. I was able to pursue a higher degree because of the selfless sacrifices you made for us – and I complete my PhD in honor of the opportunities you both deserved but were not given. Thank you for everything!

A huge thank you also to my wonderful sisters, **Briana & Carissa Amundson**, who have always been a source of inspiration, a compass in life, near-perfect role models (who often make it hard to live up to their examples!), and who are always there to remind me not to take life too seriously. They are simply the best older sisters a girl could ask for.

Family comes in many forms, including my **Chester** buddy. I adopted him when I was 20 years old and in the almost 7 years together, he has been there, right next to me (quite literally) for every long day, late night, and cross-country drive. I attribute a significant portion of my sanity through this to his constant affection and calming purrs.

Finally, but most importantly, thank you to my wonderful partner – the (newly minted) **Dr. Carl VanGessel**. Through this you've been everything – a sounding board, a listening ear, an advocate, a voice of reason, the embodiment of patience, a best friend, and so much more. Thank you for always lifting me when I was down and being there ready to celebrate every win (no matter how big or small). Your (and **Dilly's**) unconditional support has made the last few years infinitely better than they would have been without you.

DEDICATION

This dissertation is dedicated

To my parents, John & Laura Amundson, who gave me everything I needed to complete this degree over the past five years and taught me everything I needed to know in the twenty-two before that.

TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Hydraulic fracturing of subsurface shales is an important economic resource

Subsurface shale formations underly much of North America and are economically important for the recovery of oil & natural gas in the United States and globally (**Figure 1.1**)[1]. These shale formations are located deep in the subsurface and contain trapped hydrocarbons originating from organic material deposited up to 480 million years ago[2–4]. As time passed, these deposits were buried deep within the earth forming sedimentary rock while high temperatures converted the organic material to oil & gas[3,5]. Due to the low permeability of shale rock, the hydrocarbons have been stored this way for millions of years and largely remained an untapped energy resource until the late 1900s.

Development in technologies related to drilling and oil & gas recovery led to the advent of unconventional drilling techniques such as hydraulic fracturing (or 'fracking') that has facilitated access to shale energy reserves[6,7]. During hydraulic fracturing, wells are first drilled vertically and then horizontally through the target shale formation for up to 3 miles in length[8]. The 'fracturing' process involves pumping water (>90%), proppant (>5%), and chemicals (~1%) at high pressure – causing the subsurface rock formation to break and fracture (**Figure 1.1**)[9,10]. This process releases trapped hydrocarbons allowing them to flow back to the surface with water as 'produced fluids' where they are separated and recovered for refinery[8]. Without the extra steps of horizontal drilling and fracturing, the hydrocarbons would not readily diffuse out of the low-permeability rock matrix at economically recoverable rates (hence, 'unconventional').

Large-scale recovery of natural gas from shale formations has continued to grow in the United States since the first profitable shale wells were drilled in the Barnett shale (TX, USA) in

the early 2000s[6]. Since then, many companies have leveraged this technology to liberate hydrocarbon reserves in shale formations across much of the United States, as well as globally[11]. In 2016, natural gas surpassed coal-derived energy to become the primary source of energy in the United States. And as of 2022, approximately 86% of natural gas recovered in the U.S. came from shale rock[1]. Shale formations are organized within basins across the North America, Asia, Europe, and South America – with the United States, China, Canada, and Argentina estimated to have some of the largest reserves[12]. Some of the most economically relevant shale basins in the United States include the Appalachian, Permian, Bakken, Denver-Julesburg, and Western Gulf basins[12].

1.2 Characteristics of subsurface shales

Black shales are organic rich (generally 2-10% total organic content), sedimentary rocks dominated by fine-grained silt and clay minerals that are located generally between 300-4000 meters deep in the subsurface[2–4]. They are 'tight' rocks, generally exhibiting relatively low porosities that contain very little meteoric water and pore spaces (generally <0.2 μm) that tend to be very disconnected[3,13–15]. Formation temperature and pressure can varies through formations and across basins (generally between 50-125°C and 4000-10000 psi)[16,17]. The shale rock matrix is also extremely heterogenous between and within basins and even within and across an individual formation. Notably, shales vary in mineral content which directly influences the salinity of produced water from the wells[8,18]. Injected freshwater causes minerals to leach from the shale rock, converting the fresh water to saline water relatively quickly. As a result, the salinity of produced waters varies greatly between basins[16,17]. For example, produced waters from some wells in the Appalachian and Bakken basins reach extremely high salinities

(>350,000 ppm TDS), while produced waters from other basins are more analogous to sea water, such as the Denver-Julesburg and Anadarko Basin (~50,000 ppm TDS)[16,17,19].

Unfractured subsurface shale formations are an excellent environment for harboring trapped hydrocarbons, but a poor environment for microbial life. To date, no study has conclusively shown or provided strong evidence for active, native shale microorganisms. Instead, it is likely that the persisting microorganisms are introduced into the shale formation during the drilling, fracturing, and development of the well. Microbial life in pristine, unfractured shales is unlikely as the average pore space tends to be smaller than an average bacterial cell[13,15]. This, combined with lack of water and access to other nutrients, makes it an unlikely for abundant microbial activity in this environment prior to fracturing[3]. In contrast, equipment and additives used in the development of the well are not sterile and harbor microorganisms – making it probable that transfer and reuse of machinery as well the inputs and additives used (such as source water and drill muds) are responsible for the introduction of microbes into subsurface and seeding the microbiomes of fractured shale wells[20,21].

1.3 Fractured shale microbiomes

Microorganisms are ubiquitous in fractured shale systems. A subset of the microorganisms introduced into this newly formed ecosystem are able to survive and colonize despite harsh environmental conditions, resulting in a relatively low diversity microbial community that persists throughout the lifetime of the well[3]. The persisting microorganisms must be anaerobic, halotolerant, halophilic, or thermophilic to survive in the fractured shale ecosystem. The taxonomic composition of microbial communities have been described from several shale basins, including the Appalachian[3,22–35], Permian[36,37], Bakken[38–40], Denver-

Julesburg[40–42], Michigan[43,44], Illinois[45], and others[20,21,46–52]. Though some common taxa have emerged, such as *Halanaerobium* and *Thermotoga*, there is variability in the diversity and composition of shale microbiomes across basins. This is likely a direct result of strong selection pressures by temperature and salinity on the persisting microbiome, and how the strength of these pressures differ across shale basins[3]. For example, one well in the highly-saline Appalachian Basin rapidly decreased in microbial alpha diversity over time[27,28], while alpha diversity was temporally stochastic for other wells in the less-saline Anadarko Basin[21]. This likely reflects an increased metabolic and taxonomic microbial diversity in basins with lower-salinity produced fluids. Though several basins have been taxonomically profiled, there remains a need to understand functional potential and diversity of metabolisms and key traits in addition to taxonomic diversity across fractured shale systems, especially across basins that vary significantly in salinity and temperature.

Salinity presents a significant challenge to microbial life[53]. Cells must maintain positive turgor in order to retain shape, structure, and allow cellular machinery to function normally. Under elevated salinity, water from within the cytoplasm would freely diffuse through the cell membrane and into the more saline environment, causing the cell to collapse. Microorganisms have two fundamentally different strategies for counteracting elevated salinity in the environment and maintaining cell structure: (1) the 'salt-in' strategy of accumulating inorganic ions (K^+, Na^+ and Cl^-) internally and (2) biosynthesis and/or accumulation from the envrioment of small organic compounds ('compatible solutes')[54,55]. Both strategies have energetic and adaptive trade-offs. Microorganisms that use the salt-in strategy must have evolved specialized enzymes that can function despite high internal salt concentrations. However, this strategy does not require the cell expend large amounts of energy while accumulating salts internally to

maintain osmotic balance. These microorganisms are highly adapted to saline environments and often can not live in the absence of salt ('halophiles'). In contrast, microorganisms that deploy the compatible solute strategy may more readily adapt to changing environmental conditions but must use energy to synthesize or accumulate organic solutes to maintain osmotic balance[54]. These microorganisms expend more energy from their dissimilatory processes in order to survive high salt concentrations. Common compatible solutes are amino acid derivatives (i.e. glycine betaine, choline, ectoine, etc.), sugars (i.e. trehalose, sucrose, etc.) and glycerol[54]. Notably, additives used in hydraulic fracturing may be a source of these osmoprotectants that could be taken up from the environment by persisting microorganism. For example, choline could possibly be derived from choline chloride which is a common clay stabilizer used in to enhance hydrocarbon recovery[9].

Microbial metabolisms that yield enough energy to sustain life under anoxic conditions and at high salt concentrations are limited[56]. In general, the salt-in strategy is phylogenetically conserved due to high specificity for saline environments and low adaptability to other, lower saline environments. Certain lineages such as members within the order Halanaerobiales have been widely reported to use this strategy[54,55], which may partially explain the dominance of these taxa in high salinity shales such as the Appalachian, Bakken, and Permian basins (though genes for uptake of compatible solutes have also been reported in genomes of these taxa[28,30]). Additionally, certain metabolisms simply do not yield enough energy for the cell to synthesize or accumulate compatible solutes to counter osmotic stress and support life in highly saline environments[56]. For example, dissimilatory sulfate reduction, acetoclastic and hydrogenotrophic methanogenesis are absent from ecosystems with elevated salinities for this reason – with sulfate reduction generally limited at 120-200 g/L salts and acetoclastic and hydrogenotrophic

methanogenesis limited even to even lower saline environments, between 30-100 g/L salts[56]. This in part explains the apparent lack of anaerobic respiration with sulfate in some basins exhibiting high salinity produced fluids, such as the Appalachian basin[28]. Overall, microbial metabolisms dominant in fractured shale ecosystems are heavily influenced, and more so limited by, the trade-off between energy gained from a given metabolism, the salt-tolerance strategy of a microorganism, and the energy required balance osmotic stress driven by elevated salinities.

Viruses have also been reported in fractured shales and likely play an important role in nutrient cycling in these closed ecosystems[27–30]. Though very small in size, even relative to their bacterial and archaeal hosts, viruses can have a profound impact on the microbial community through top-down and bottom-up controls[27]. For example, the impact of viral predation has been well illustrated in oceanic ecosystems where a majority of bacteria are estimated to be lysed daily, releasing important nutrients and impacting biogeochemical cycling[57]. Top-down control by viruses on the microbial community results in lysing of dominant microbial host populations, which can have an impact on the functioning of the shale microbiome. However, in turn, cell lysis also releases important metabolites. This bottom-up control is like very important in closed ecosystems such as fractured shales and could help to sustain microbial life for longer periods of time as metabolites released may be used by other remaining microbial community members. In fractured shale systems, cell lysis could release important organic compounds such as osmoprotectants or other substrates that may support metabolisms, such as fermentation[27,28,30]. In general, it is likely that both controls significantly contribute to fractured shale microbial community dynamics and persistence through time.

Viruses are the only known predator within fractured shales. In response, many hosts encode viral defense systems, such as adaptive immunity through CRISPR-Cas, which is likely

an important trait contributing to microbial persistence over extended periods of time[42,58–62]. The interactions between hosts and viruses are often challenging to detangle in many ecosystems but are crucial to understanding temporal dynamics of microbial communities especially in fractured shales[63–65]. Spacers within CRISPR-Cas arrays allow the opportunity to make host-virus linkages and better understand the dynamics between hosts and viruses, and what effect that may have on the overall microbial community through top down or bottom up controls.

Persisting shale microorganisms contribute to many common production challenges, such as biofilm formation, generation of organic acids and toxic hydrogen sulfide[66–71]. Biofilm formation by microorganisms can contribute to biologging, or clogging of the shale factures, fissures, and pores that allow for hydrocarbons to diffuse out of the rock matrix and flow to the surface[24,31]. However, biofilms allow microorganisms to attach to the rock face and sustain shearing pressures and may provide resistance against biocide treatment[72]. Biofilms also facilitate metabolic exchange which may ultimately facilitate microbial survival in the harsh shale ecosystem[72,73]. Biological sulfidogenesis, or generation of hydrogen sulfide, is a direct byproduct of respiratory sulfate reduction and thiosulfate reduction. Sulfate and thiosulfate, likely originating from the shale rock[8], can be reduced by some microorganisms to generate energy that sustains microbial life in saline and anaerobic envrioments[54,56]. However, hydrogen sulfide is toxic and can contribute to 'souring' of the well, in addition to scaling – or precipitation of sulfide minerals that accumulate on well infrastructure[32,38,69]. Finally, production of organic acids also can contribute to corrosion of well infrastructure, especially at high concentrations[74,75]. However, fermentation is a dominant metabolism considering anaerobic and reduced redox conditions of fractured shales and has been reported across shale basins, and thus may be a persistent problem in fractured shales[21,28].

Industrial operators invest large amounts of time and money into biocide applications within wells and associated top-side infrastructure in response to the many possible deleterious impacts caused by persisting microorganisms[9,71]. However, many of these biocides may not be chemically active within the subsurface at high temperatures and pressures – ultimately rendering them functionally useless[17,70,71]. In this way persisting shale microbiomes are directly impacted by their environment (high salinity, temperature, biocides etc.), yet also impact their environment in response (biofilm formation, sulfide generation, etc.).

1.4 Methods for studying shale microbiomes

Historical approaches for studying microorganisms in oil and gas systems have focused primarily on methods that do not fully capture the functional potential of the persisting microbiomes[76]. Traditional culture-based methods are a common approach to studying microbial function, however the 'great plate anomaly' tells us that most microbes cannot be cultured under laboratory conditions, and if they can they likely will not exhibit the same phenotypes and physiological traits as they would *in situ*[77]. Another common approach for understanding the functional potential of a microbial community is PCR amplification of highly characterized functional genes. An example of this is in oil and gas systems is amplification of the sulfite reductase (*dsrB*) gene, which is a well characterized marker gene for sulfate reduction and thus an indicator for production of toxic hydrogen sulfide[78,79]. However, a major flaw in this approach is that (1) there often exists several pathways for end products of interest, such as rhodanese or thiosulfate reductase (*phsA*) genes which reduce thiosulfate instead of sulfate but also result in production of hydrogen sulfide, and (2) designing successful primers for marker genes can be challenging as many genes are not well evolutionarily conserved or may be poorly annotated.

Finally, with the advent of genomic sequencing, 16S rRNA marker gene sequencing has also become a common approach to taxonomically profiling microbial communities[80]. However, this type of sequencing lacks species and strain-level resolution and does not provide insights into functional potential of the microbial community.

Metagenomic sequencing offers the opportunity to unravel the functional potential of shale microbiomes more comprehensively. By sequencing all the DNA from a sample and assembling it into metagenome assembled genomes (MAGs), functional potential can be evaluated at a genome-resolved level (**Figure 1.2**)[81]. This enables insights into the microbial community down to a strain-level resolution of taxonomic classification while also gaining insights into what genes are encoded within these populations. Thus, metagenomics allows us to move beyond simply gene-centric or taxonomy-centric views of microbial communities and instead discern both together for individual microbial populations. However, one challenge that remains with a metagenomic approach is that low abundance microorganisms may be easily missed with insufficient depth of sequencing data, but building large metagenomic datasets to fully capture the microbial community quickly becomes very computationally expensive.

1.5 A metagenomic approach for investigating shale microbiomes

A myriad of questions remain about microbial life in fractured shales. The vast majority of research on microbial communities in these systems has been focused on profiling community composition using 16S rRNA amplicon sequencing. Furthermore, much of this research has been conducted on samples from the Appalachian Basin, while many other relevant basins remain far less characterized. While the body of existing research is important, there remains a crucial need

to study the functional potential of shale microbiomes in addition to taxonomic composition, and to do this across a diversity of shale basins.

My dissertation research is focused on using metagenomic sequencing and paired metabolomics to study the functional potential of shale microbiomes from many different shale basins across North America. Specifically, I used this approach to better understand the mechanisms and traits that contribute to colonization and survival of microorganisms across heterogenous shale formations. In total, this dissertation encompasses research on 209 metagenomes across 11 shale basins, resulting in over 4.2 terabases of sequencing.

The first aim of this work (**Chapter 2**) was to detangle the key microbial players and their functional potentials from a western shale basin that exhibited lower salinities and higher temperatures than the well-characterized Appalachian Basin[21]. At the time of this publication, no studies using metagenomics had been published on the shale microbiomes outside of the Appalachian Basin. This research shed light on the potential for dissimilatory sulfate reduction from metagenomic derived evidence, highlighted the potential for microorganisms to use chemical inputs as a source of energy, and provided evidence for the source of microorganisms in fractured shales through a topside audit and source tracking analysis.

The second aim of this work (**Chapter 3**) was to better understand the role of CRISPR-Cas systems in host defense against viruses through a longitudinal study[42]. Viral defense systems, such as CRISPR, are another important trait for persisting microorganisms as viruses are the only bacterial and archaeal predator in fractured shale ecosystems. This study highlighted the ongoing arms-race and coexistence that likely occurs between hosts and viruses in the subsurface and other closed ecosystems.

Finally, the third aim of this work (**Chapter 4**) was to holistically evaluate metabolisms and other key traits of colonizing and persisting microbiomes at a broader, basin-scale. This research leveraged metagenomic sequencing from all 11 shale basins to build a database and tool dedicated to studying shale microbiomes. The results from this research chapter have the potential to inform microbial management practices within fractured shale systems across North America and elsewhere.

Ultimately, the sum of this research contributes a new and substantial insights into economically important fractured shale systems and highlights a metagenomic approach to understanding how microbiomes are effected by their environment, and how emergent properties of microbiomes can also effect the ecosystem.

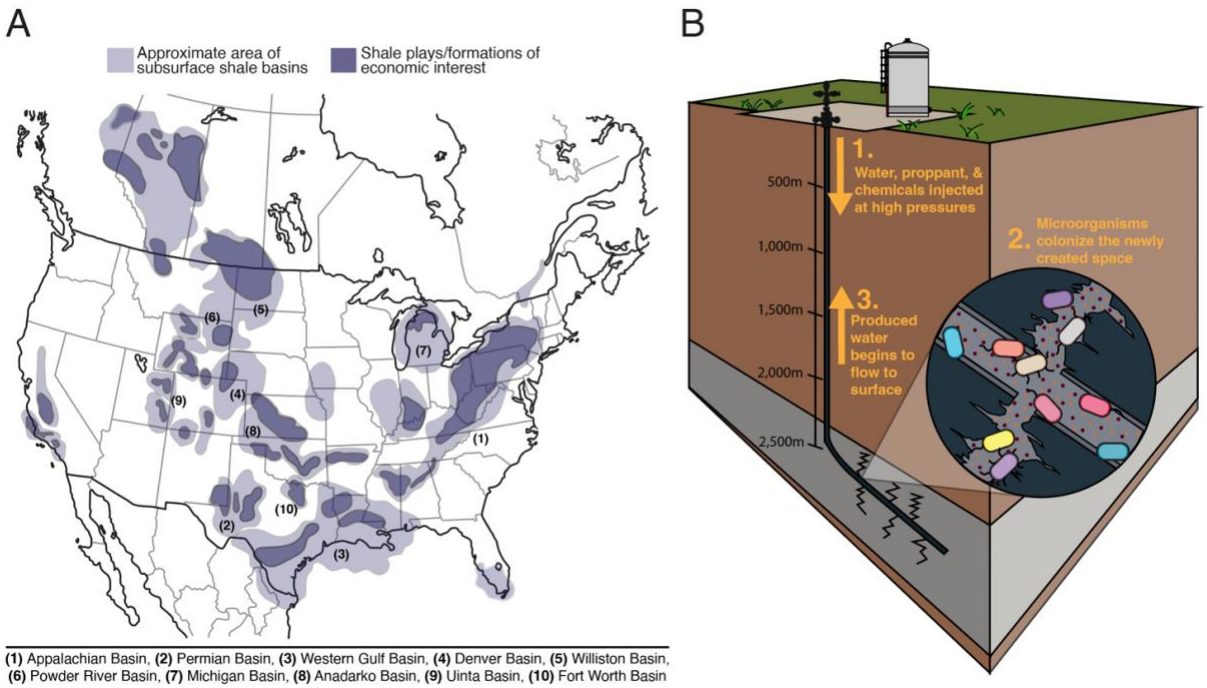

Figure 1.1. (A) Distribution of subsurface shales in North America. The approximate location of shale basins are outlined in light purple, with dark purple indicating shale plays or formations that are a target for oil & gas recovery. The top ten basins estimated to have the greatest remaining resources (as of 2022) are labeled in descending order. (B) Schematic depicting the process of hydraulic fracturing of shale formations and microbial colonization of this system.

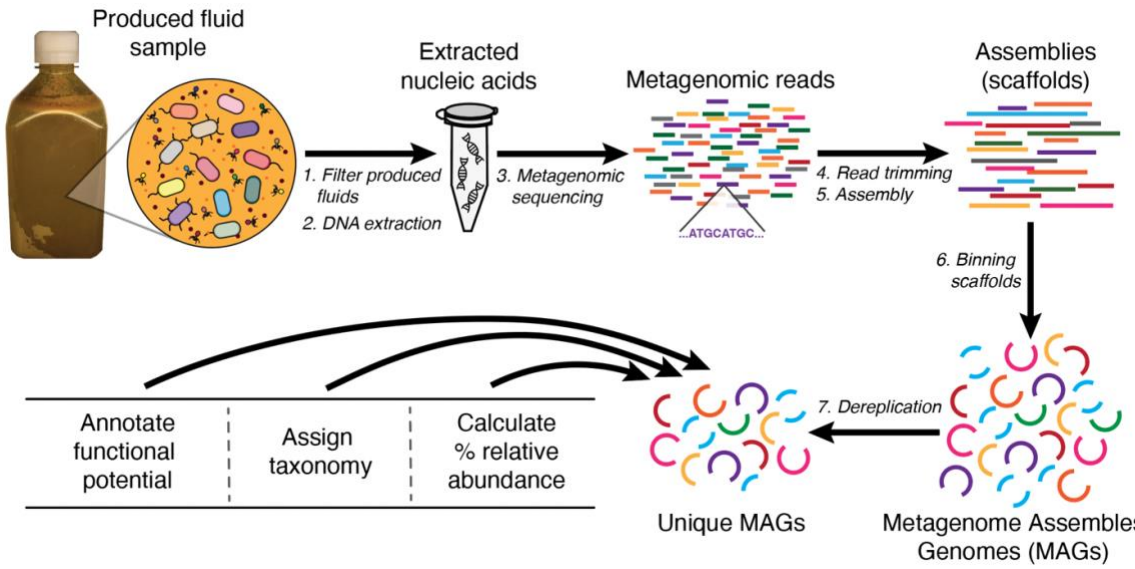

Figure 1.2. Workflow of processing a raw produced fluid sample for metagenomic sequencing and processing the raw sequencing data for a genome-resolved approach to studying shale microbiomes.

Chapter 1 References

1. Where our natural gas comes from - U.S. Energy Information Administration (EIA). https://www.eia.gov/energyexplained/natural-gas/where-our-natural-gas-comes-from.php.
2. Tourtelot, H. A. Black Shale—Its Deposition and Diagenesis. *Clays Clay Miner.* **27**, 313–321 (1979).
3. Mouser, P. J., Borton, M., Darrah, T. H., Hartsock, A. & Wrighton, K. C. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol. Ecol.* **92**, fiw166 (2016).
4. Wang, Q., Chen, X., Jha, A. N. & Rogers, H. Natural gas from shale formation – The evolution, evidences and challenges of shale gas revolution in United States. *Renew. Sustain. Energy Rev.* **30**, 1–28 (2014).
5. Trabucho-Alexandre, J., Hay, W. W. & de Boer, P. L. Phanerozoic environments of black shale deposition and the Wilson Cycle. *Solid Earth* **3**, 29–42 (2012).
6. Kerr, R. A. Natural Gas From Shale Bursts Onto the Scene. *Science* **328**, 1624–1626 (2010).
7. Tour, J. M., Kittrell, C. & Colvin, V. L. Green carbon as a bridge to renewable energy. *Nat. Mater.* **9**, 871–874 (2010).
8. Khan, H. J. *et al.* A Critical Review of the Physicochemical Impacts of Water Chemistry on Shale in Hydraulic Fracturing Systems. *Environ. Sci. Technol.* **55**, 1377–1394 (2021).
9. Ferrer, I. & Thurman, E. M. Chemical constituents and analytical approaches for hydraulic fracturing waters. *Trends Environ. Anal. Chem.* **5**, 18–25 (2015).
10. Sun, Y. *et al.* A critical review of risks, characteristics, and treatment strategies for potentially toxic elements in wastewater from shale gas extraction. *Environ. Int.* **125**, 452–469 (2019).
11. U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. *World Shale Resource Assessments* https://www.eia.gov/analysis/studies/worldshalegas/.
12. Assumptions to the Annual Energy Outlook 2023: Oil and Gas Supply Module. (2023).
13. Fredrickson, J. K. *et al.* Pore-size constraints on the activity and survival of subsurface bacteria in a late cretaceous shale-sandstone sequence, northwestern New Mexico. *Geomicrobiol. J.* **14**, 183–202 (1997).

14. Soeder, D. J. Porosity and Permeability of Eastern Devonian Gas Shale. *SPE Form. Eval.* **3**, 116–124 (1988).
15. Onstott, T. C. *et al.* Observations pertaining to the origin and ecology of microorganisms recovered from the deep subsurface of Taylorsville Basin, Virginia. *Geomicrobiol. J.* **15**, 353–385 (1998).
16. Shaffer, D. L. *et al.* Desalination and Reuse of High-Salinity Shale Gas Produced Water: Drivers, Technologies, and Future Directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
17. Kahrilas, G. A., Blotevogel, J., Corrin, E. R. & Borch, T. Downhole Transformation of the Hydraulic Fracturing Fluid Biocide Glutaraldehyde: Implications for Flowback and Produced Water Quality. *Environ. Sci. Technol.* **50**, 11414–11423 (2016).
18. Wood, D. A. & Hazra, B. Characterization of organic-rich shales for petroleum exploration & exploitation: A review-Part 1: Bulk properties, multi-scale geometry and gas adsorption. *J. Earth Sci.* **28**, 739–757 (2017).
19. Blondes, M. S. *et al.* U.S. Geological Survey National Produced Waters Geochemical Database v2.3. (2019) doi:10.5066/F7J964W8.
20. Struchtemeyer, C. G., Davis, J. P. & Elshahed, M. S. Influence of the Drilling Mud Formulation Process on the Bacterial Communities in Thermogenic Natural Gas Wells of the Barnett Shale. *Appl. Environ. Microbiol.* **77**, 4744–4753 (2011).
21. Amundson, K. K. *et al.* Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. *Microbiome* **10**, 5 (2022).
22. Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K. & Mouser, P. J. Temporal Changes in Microbial Ecology and Geochemistry in Produced Water from Hydraulically Fractured Marcellus Shale Gas Wells. *Environ. Sci. Technol.* **48**, 6508–6517 (2014).
23. Lipus, D. *et al.* Predominance and Metabolic Potential of Halanaerobium spp. in Produced Water from Hydraulically Fractured Marcellus Shale Wells. *Appl. Environ. Microbiol.* **83**, e02659-16 (2017).
24. Vikram, A., Lipus, D. & Bibby, K. Metatranscriptome analysis of active microbial communities in produced water samples from the Marcellus Shale. *Microb. Ecol.* **72**, 571–581 (2016).

25. Evans, M. V. *et al.* Members of Marinobacter and Arcobacter Influence System Biogeochemistry During Early Production of Hydraulically Fractured Natural Gas Wells in the Appalachian Basin. *Front. Microbiol.* **9**, (2018).
26. Methanogenic Archaea in Marcellus Shale: A Possible Mechanism for Enhanced Gas Recovery in Unconventional Shale Resources | Environmental Science & Technology. https://pubs.acs.org/doi/full/10.1021/acs.est.5b00765.
27. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
28. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).
29. Borton, M. A. *et al.* Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale. *Environ. Microbiol.* **20**, 4596–4611 (2018).
30. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci.* **115**, E6585–E6594 (2018).
31. Booker, A. E. *et al.* Deep-Subsurface Pressure Stimulates Metabolic Plasticity in Shale-Colonizing Halanaerobium spp. *Appl. Environ. Microbiol.* **85**, e00018-19 (2019).
32. Booker, A. E. *et al.* Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* **2**, 10.1128/mspheredirect.00257-17 (2017).
33. Murali Mohan, A., Hartsock, A., Hammack, R. W., Vidic, R. D. & Gregory, K. B. Microbial communities in flowback water impoundments from hydraulic fracturing for recovery of shale gas. *FEMS Microbiol. Ecol.* **86**, 567–580 (2013).
34. Murali Mohan, A. *et al.* Microbial Community Changes in Hydraulic Fracturing Fluids and Produced Water from Shale Gas Extraction. *Environ. Sci. Technol.* **47**, 13141–13150 (2013).
35. Akob, D. M., Cozzarelli, I. M., Dunlap, D. S., Rowan, E. L. & Lorah, M. M. Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Appl. Geochem.* **60**, 116–125 (2015).
36. Tinker, K., Lipus, D., Gardiner, J., Stuckman, M. & Gulliver, D. The Microbial Community and Functional Potential in the Midland Basin Reveal a Community

Dominated by Both Thiosulfate and Sulfate-Reducing Microorganisms. *Microbiol. Spectr.* **10**, e00049-22 (2022).

37. Kilbane, J., Wylde, J. & Williamson, A. Investigation of Microorganisms in a West Texas Oilfield Using Growth and Genetic Testing. in (OnePetro, 2015). doi:10.2118/173709-MS.
38. An, B. A., Shen, Y. & Voordouw, G. Control of Sulfide Production in High Salinity Bakken Shale Oil Reservoirs by Halophilic Bacteria Reducing Nitrate to Nitrite. *Front. Microbiol.* **8**, (2017).
39. Tinker, K. *et al.* Geochemistry and Microbiology Predict Environmental Niches With Conditions Favoring Potential Microbial Activity in the Bakken Shale. *Front. Microbiol.* **11**, (2020).
40. Wang, H., Lu, L., Chen, X., Bian, Y. & Ren, Z. J. Geochemical and microbial characterizations of flowback and produced water in three shale oil and gas plays in the central and western United States. *Water Res.* **164**, 114942 (2019).
41. Hull, N. M., Rosenblum, J. S., Robertson, C. E., Harris, J. K. & Linden, K. G. Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. *Sci. Total Environ.* **644**, 183–192 (2018).
42. Amundson, K. K., Roux, S., Shelton, J. L. & Wilkins, M. J. Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales. *Curr. Biol.* **33**, 3125-3135.e4 (2023).
43. Kirk, M. F. *et al.* Impact of commercial natural gas production on geochemistry and microbiology in a shale-gas reservoir. *Chem. Geol.* **332–333**, 15–25 (2012).
44. Stemple, B. *et al.* Biogeochemistry of the Antrim Shale Natural Gas Reservoir. *ACS Earth Space Chem.* **5**, 1752–1761 (2021).
45. Schlegel, M. E., McIntosh, J. C., Bates, B. L., Kirk, M. F. & Martini, A. M. Comparison of fluid geochemistry and microbiology of multiple organic-rich reservoirs in the Illinois Basin, USA: Evidence for controls on methanogenesis and microbial transport. *Geochim. Cosmochim. Acta* **75**, 1903–1919 (2011).
46. Zhang, Y., Yu, Z., Zhang, H. & Thompson, I. P. Microbial distribution and variation in produced water from separators to storage tanks of shale gas wells in Sichuan Basin, China. *Environ. Sci. Water Res. Technol.* **3**, 340–351 (2017).

47. Hernandez-Becerra, N. *et al.* New microbiological insights from the Bowland shale highlight heterogeneity of the hydraulically fractured shale microbiome. *Environ. Microbiome* **18**, 14 (2023).
48. Zhong, C. *et al.* Temporal Changes in Microbial Community Composition and Geochemistry in Flowback and Produced Water from the Duvernay Formation. *ACS Earth Space Chem.* **3**, 1047–1057 (2019).
49. Wu, H. *et al.* Temporal changes of bacterial and archaeal community structure and their corrosion mechanisms in flowback and produced water from shale gas well. *J. Nat. Gas Sci. Eng.* **104**, 104663 (2022).
50. Santillan, E.-F. U., Choi, W., Bennett, P. C. & Diouma Leyris, J. The effects of biocide use on the microbiology and geochemistry of produced water in the Eagle Ford formation, Texas, U.S.A. *J. Pet. Sci. Eng.* **135**, 1–9 (2015).
51. Mu, H. M. *et al.* A rapid change in microbial communities of the shale gas drilling fluid from 3548 m depth to the above-ground storage tank. *Sci. Total Environ.* **784**, 147009 (2021).
52. Struchtemeyer, C. G. & Elshahed, M. S. Bacterial communities associated with hydraulic fracturing fluids in thermogenic natural gas wells in North Central Texas, USA. *FEMS Microbiol. Ecol.* **81**, 13–25 (2012).
53. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci.* **104**, 11436–11440 (2007).
54. Oren, A. Life at High Salt Concentrations. in *The Prokaryotes: Prokaryotic Communities and Ecophysiology* (eds. Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 421–440 (Springer, 2013). doi:10.1007/978-3-642-30123-0_57.
55. Oren, A. Bioenergetic Aspects of Halophilism. *Microbiol. Mol. Biol. Rev.* **63**, 334–348 (1999).
56. Oren, A. Thermodynamic limits to microbial life at high salt concentrations. *Environ. Microbiol.* **13**, 1908–1923 (2011).
57. Chow, C.-E. T. & Suttle, C. A. Biogeography of Viruses in the Sea. *Annu. Rev. Virol.* **2**, 41–66 (2015).
58. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

59. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).
60. Staals, R. H. J. & Brouns, S. J. J. Distribution and Mechanism of the Type I CRISPR-Cas Systems. in *CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea* (eds. Barrangou, R. & van der Oost, J.) 145–169 (Springer, 2013). doi:10.1007/978-3-642-34657-6_6.
61. Berg, M. *et al.* Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. *ISME J.* **15**, 1569–1584 (2021).
62. Guerrero, L. D. *et al.* Long-run bacteria-phage coexistence dynamics under natural habitat conditions in an environmental biotechnology system. *ISME J.* **15**, 636–648 (2021).
63. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
64. Anderson, R. E., Brazelton, W. J. & Baross, J. A. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol. Ecol.* **77**, 120–133 (2011).
65. Sanguino, L., Franqueville, L., Vogel, T. M. & Larose, C. Linking environmental prokaryotic viruses and their host through CRISPRs. *FEMS Microbiol. Ecol.* **91**, fiv046 (2015).
66. Gaspar, J. *et al.* Microbial Dynamics and Control in Shale Gas Production. *Environ. Sci. Technol. Lett.* **1**, 465–473 (2014).
67. Bakke, R., Rivedal, B. & Mehan, S. Oil reservoir biofouling control. *Biofouling* **6**, 53–60 (1992).
68. Jew, A. D. *et al.* Impact of Organics and Carbonates on the Oxidation and Precipitation of Iron during Hydraulic Fracturing of Shale. *Energy Fuels* **31**, 3643–3658 (2017).
69. Gieg, L. M., Jack, T. R. & Foght, J. M. Biological souring and mitigation in oil reservoirs. *Appl. Microbiol. Biotechnol.* **92**, 263–282 (2011).
70. Johnson, K., French, K., Fichter, J. K. & Oden, R. Use Of Microbiocides In Barnett Shale Gas Well Fracturing Fluids To Control Bacteria Related Problems. in (OnePetro, 2008).

71. Struchtemeyer, C. G., Morrison, M. D. & Elshahed, M. S. A critical assessment of the efficacy of biocides used during the hydraulic fracturing process in shale natural gas wells. *Int. Biodeterior. Biodegrad.* **71**, 15–21 (2012).
72. Bamford, N. C., MacPhee, C. E. & Stanley-Wall, N. R. Microbial Primer: An introduction to biofilms – what they are, why they form and their impact on built and natural environments. *Microbiology* **169**, 001338 (2023).
73. Zhou, L. *et al.* Rates of Sulfate Reduction by Microbial Biofilms That Form on Shale Fracture Walls within a Microfluidic Testbed. *ACS EST Eng.* **2**, 1619–1631 (2022).
74. Liang, R. *et al.* Metabolic Capability of a Predominant Halanaerobium sp. in Hydraulically Fractured Gas Wells and Its Implication in Pipeline Corrosion. *Front. Microbiol.* **7**, (2016).
75. Alabbas, F. M. & Mishra, B. Microbiologically Influenced Corrosion of Pipelines in the Oil & Gas Industry. in *Proceedings of the 8th Pacific Rim International Congress on Advanced Materials and Processing* (ed. Marquis, F.) 3441–3448 (Springer International Publishing, 2016). doi:10.1007/978-3-319-48764-9_426.
76. Bhagobaty, R. K. Culture dependent methods for enumeration of sulphate reducing bacteria (SRB) in the Oil and Gas industry. *Rev. Environ. Sci. Biotechnol.* **13**, 11–16 (2014).
77. Staley, J. T. & Konopka, A. Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu. Rev. Microbiol.* **39**, 321–346 (1985).
78. Geets, J. *et al.* DsrB gene-based DGGE for community and diversity surveys of sulfate-reducing bacteria. *J. Microbiol. Methods* **66**, 194–205 (2006).
79. Mohanakrishnan, J. *et al.* Dynamic microbial response of sulfidogenic wastewater biofilm to nitrate. *Appl. Microbiol. Biotechnol.* **91**, 1647–1657 (2011).
80. Kotu, S. P., Mannan, M. S. & Jayaraman, A. Emerging molecular techniques for studying microbial community composition and function in microbiologically influenced corrosion. *Int. Biodeterior. Biodegrad.* **144**, 104722 (2019).
81. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

Chapter 2: Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation[1]

2.1 Summary

Microbial colonization of subsurface shales following hydraulic fracturing offers the opportunity to study coupled biotic and abiotic factors that impact microbial persistence in engineered deep subsurface ecosystems. Shale formations underly much of the continental USA and display geographically distinct gradients in temperature and salinity. Complementing studies performed in eastern USA shales that contain brine-like fluids, here we coupled metagenomic and metabolomic approaches to develop the first genome-level insights into ecosystem colonization and microbial community interactions in a lower-salinity, but high-temperature western USA shale formation. We collected materials used during the hydraulic fracturing process (i.e., chemicals, drill muds) paired with temporal sampling of water produced from three different hydraulically fractured wells in the STACK (Sooner Trend Anadarko Basin, Canadian and Kingfisher) shale play in OK, USA. Relative to other shale formations, our metagenomic and metabolomic analyses revealed an expanded taxonomic and metabolic diversity of microorganisms that colonize and persist in fractured shales. Importantly, temporal sampling across all three hydraulic fracturing wells traced the degradation of complex polymers from the hydraulic fracturing process to production and consumption of organic acids that support sulfate- and thiosulfate-reducing bacteria. Furthermore, we identified 5,587 viral genomes and linked many of these to the dominant, colonizing microorganisms, demonstrating the key role that viral

[1] This chapter was reproduced verbatim from "Amundson, et al. Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. Microbiome J (2022)". The text benefitted from writing and editing contributions from contributing authors and reviewers selected by the publisher. The ordering of the materials in this dissertation are consistent with the content available online but have been renumbered to reflect incorporation into this dissertation.

predation plays in community dynamics within this closed, engineered system. Lastly, top-side audit sampling of different source materials enabled genome-resolved source tracking, revealing the likely sources of many key colonizing and persisting taxa in these ecosystems. These findings highlight the importance of resource utilization and resistance to viral predation as key traits that enable specific microbial taxa to persist across fractured shale ecosystems. We also demonstrate the importance of materials used in the hydraulic fracturing process as both a source of persisting shale microorganisms and organic substrates that likely aid in sustaining the microbial community. Moreover, we showed that different physicochemical conditions (i.e., salinity, temperature) can influence the composition and functional potential of persisting microbial communities in shale ecosystems. Together, these results expand our knowledge of microbial life in deep subsurface shales and have important ramifications for management and treatment of microbial biomass in hydraulically fractured wells.

2.2 Introduction

Deep terrestrial shale formations underly much of North America, and due to extremely low permeability and very small, disconnected pore spaces (~10 nm)[1] are generally thought to lack suitable habitat for microbial life[2]. However, the high-pressure injection of water, sand, and chemicals into deep shales as part of the hydraulic fracturing (HF) process transform this environment, resulting in formation of extensive fracture networks within the rock matrix[3]. Microorganisms present in injected materials (e.g., drill muds) colonize these new fracture networks and encounter nutrient rich HF additives that act as substrates for microbial growth[4–6]. Under these conditions, established microbial communities can persist for extended periods of time (>300 days)[5,7].

Although these fractured shales differ from pristine subsurface ecosystems, they are confined by the surrounding hard rock matrix and are isolated from any other sources of microbial immigration[2,8]. As such, they can be leveraged to investigate the metabolic strategies and interactions that govern microbial community dynamics in engineered subsurface ecosystems. Furthermore, the process of HF to recover natural gas and oil is a critical component of the United States energy portfolio[9]. By-products of microbial metabolism such as sulfides and organic acids often drive deleterious processes, such as corrosion of infrastructure and souring of hydrocarbon streams[10–13]. Therefore, an improved understanding of microbial processes under relevant *in situ* conditions is critical to inform safer and more targeted microbial management.

To date, microbial communities have been analyzed in produced fluids from HF operations in the Appalachian Basin (Utica and Marcellus formations)[5,7,14–16], as well as the Bakken[17–19], Barnett[6,18,20], and Niobrara[18,21] shale formations. Recovered fluids from hydraulically fractured shales are often highly saline (25-200 mS/cm) but this varies greatly across geographic locations. For example, shales in the Appalachian Basin and the Bakken formation often exhibit brine-like salinities[5,17,18] but other formations, such as the STACK shale play in the Anadarko Basin, tend to display lower salinities. Indeed, glycine betaine cycling has been shown to be an important microbial process in supporting osmoprotection and energy needs of the persisting microbial community in the more saline Appalachian Basin, which is dominated by *Halanaerobium* populations[22]. The degradation of these osmoprotectants also yielded precursor compounds for methylotrophic methanogenesis[22,23], a biogenic source of natural gas.

Prior genomic investigations of microbiomes colonizing Marcellus and Utica formations by our group also highlighted the key role of viruses in fractured shale ecosystems. In these wells, viral predation was inferred to contribute to strain-level dynamics in dominant

Halanaerobium populations and catalyze the release of labile cellular compounds via cell lysis to support the persisting community members[24]. Finally, elevated salinity in many of these systems can inhibit the growth of canonical sulfate reducing microorganisms[25–27], with thiosulfate-dependent sulfidogenesis catalyzed by rhodanese enzymes instead identified as the dominant pathway for sulfide production[4,13]. Together, these results offered insights into fractured shale ecosystems characterized by brine-level salinity and intermediate (i.e., 50-100°C) *in situ* temperatures[28], revealing the dominance of a single species of *Halanaerobium*, importance of viral predation, and sulfide production from sources other than sulfate.

In contrast to these well-studied HF wells, many shale environments, especially those in the western and southern US, are characterized by significantly lower salinity but higher *in situ* temperatures, that can range between 100-120°C[29]. Here the STACK shale play is leveraged as an example of a western shale play characterized by differing physiochemical conditions, relative to eastern shale plays, to interrogate microbial communities. Given the specific metabolic and physiological adaptations that microorganisms encode to tolerate these physicochemical conditions[25,30,31], we hypothesized that this variability in salinity and temperature could significantly impact microbial community composition and function. Here we recovered samples of input materials and temporally sampled produced fluids from three hydraulically fractured wells in the STACK shale play (OK, USA)[28]. Through integrated metagenomic and metabolomic analyses we identified the sources of colonizing microorganisms and uncover the key metabolisms and metabolic hand-offs that enable microbial persistence in lower salinity, higher temperature engineered subsurface ecosystems.

2.3 Results & Discussion

2.3.1 Deep subsurface physicochemical conditions enrich for a conserved microbial community over time

Chemical and microbial dynamics were interrogated across three wells within the STACK shale play, OK, USA. Differences in drilling and HF techniques between the STACK-14 and STACK-16 & 17 wells (**Appendix B**) afforded a unique opportunity to investigate how variability in the chemistry and microbiology of the input fluids ('frack fluids') used in fracturing of the shale influenced the microbial community assembly over time.

Microbiological and chemical analyses revealed that the frack fluids (**Figure 2.1**) for each well (STACK-14 vs. STACK-16 & 17) had statistically discernable starting microbial communities (**Figure 2.2**) and metabolite chemistries (**Table 2.1**), as measured by 16S rRNA gene sequencing and Nuclear Magnetic Resonance (NMR) spectroscopy, respectively. For example, choline and isopropanol were discriminant chemical features in STACK-14 frack fluids, while acetate and glutarate were discriminant compounds in STACK-16 & 17 frack fluids (**Table 2.1**). However, despite initial differences in microbial community composition and chemical inputs, microbial communities in produced fluids collected 100 days after HF could no longer be statistically distinguished between the wells, suggesting deep subsurface shale conditions enriched for similar microbial taxa.

Metagenome-derived insights into community composition and dynamics mirrored observations made with complementary 16S rRNA gene datasets. Briefly, dominant taxa across both datasets were affiliated with Thermotogae, Fusobacteriales, and Clostridia. Focusing on metagenomic analyses, the dominant microbial community members between the three wells were represented by 24 metagenome assembled genomes (MAGs) (achieving >5% relative

abundance at any time point) (**Appendix B**). We observed the dominance of a single, high-quality Thermotoga petrophila MAG (M2-7-6-bin.8) (92% complete, <2% contamination) in the majority of all 18 produced fluid timepoints across the 3 wells and note the overwhelming dominance of this Thermotoga MAG through the entire STACK-17 timeseries (**Figure 2.3**). The remainder of the microbial community across the STACK wells was dominated by MAGs affiliated with Firmicutes, Desulfobacterota, and Bacteroidota, with only one Archaeal MAG recovered (Halobacterota). Three MAGs were affiliated with two novel genera, Clostridia SK-Y3 (K-7-4-bin.6) and Peptococcia DRI-13 (M1-7-4-bin.22). We were unable to assign family-level placement for two MAGs (Fusobacteriales (K-7-4-bin.55) and Desulfitibacterales (K-7-2-bin.50)), highlighting their taxonomic novelty.

Metabolic characterization of these MAGs revealed that samples were dominated by inferred fermenters and sub-populations of inferred respiratory sulfate- and thiosulfate-reducing microorganisms (**Appendix B**). Functional profiling of the prevalent Thermotoga MAG (M2-7-6-bin.8) revealed a fermentative lifestyle with the capacity for both simple and complex carbon degradation, findings similar to laboratory-based physiological studies of this genus[53–56]. Other key taxa inferred to be fermenters were affiliated with the classes Clostridia, Mahellia, and Bacteroidia. All inferred fermenters lacked genomic evidence of a complete electron transport chain, and here we cataloged the possible organic carbon sources for growth via inventorying the genes encoding carbohydrate active enzymes (CAZymes) (**Figure 2.4**). MAGs that represent taxa inferred to perform sulfur cycling were affiliated with Classes Desulfovibrionia, Deferribacteres, Syntrophobacteria, Peptococcia, and Moorellia and were characterized by the presence of reductive dsrAB and/or phsA genes, smaller complements of CAZymes and more complete electron transport chains (**Figure 2.4**). Together, these dominant microorganisms have

the potential to produce corrosive sulfide and organic acids, which are highly detrimental to the recovery of oil & gas in these systems.

2.3.2 Genome-resolved source-tracking reveals hydraulic fracturing inputs play a crucial role in the inoculation of dominant microorganisms in fractured shale ecosystems.

Given that deep shale formations are most likely devoid of microbial life prior to HF, a key goal for management of these systems is determining the source of microbial taxa that subsequently colonize and persist within the fracture network. Previous studies by our research group and others have hypothesized that exogenous microorganisms introduced during the HF process are responsible for inoculating the fracture network[5,6]. Here we leveraged a novel and extensive catalog of input samples used in the development of the STACK-14, 16, & 17 wells to perform genome-resolved source tracking of 24 dominant MAGs in support of this hypothesis.

By mapping metagenomic reads from input samples to MAGs, we detected genomic signatures for five of the 24 dominant and persisting microorganisms in input samples (**Figure 2.5**), providing the first detailed source tracking of persisting, dominant microbes during the well engineering. Not all 24 dominant MAGs had detectable signals in input materials, however this is likely due to the physical complexity of the materials and sequencing depth of samples rather than evidence of indigenous microbial life. Microorganisms that persist in hydraulically fractured shales often have metabolic potential to produce corrosive organic acids or sulfides which damage well infrastructure and interfere with oil and gas recovery. Indeed, two MAGs representing inferred fermentative taxa, including the dominant *Thermotoga* MAG (M2-7-6-bin.8), were identified in source water and frack fluids, while the SK-Y3 *Clostridia* MAG (K-7-4-bin.6) was detected in drill muds. Notably, three MAGs with putative roles in sulfur cycling

(*Shewanella*, *Peptococcia*; DRI-13, *Desulfitibacterales*) were also detected in drill muds (Figure S3). The detection of four out of five key MAGs in the drill muds suggests that these organic-rich materials likely harbor key taxa that colonize the fracture network[57]. As such, these materials may require more targeted microbial control practices to minimize subsurface biomass growth. Additionally, the detection of the dominant *Thermotoga* genome in frack fluids offers strong evidence that this microorganism is derived from surface inputs. Given the prevalence of microorganisms in fractured shale ecosystems and the consequences of their metabolic by-products on subsurface infrastructure and resources, understanding these sources of biomass is crucial for targeted microbial management.

2.3.3 *Organic additives used in the hydraulic fracturing process are a nutrient resource for shale colonizing microbial members*

Complex organic additives used during the HF process may be degraded by colonizing microorganisms, potentially yielding more labile substrates[58]. To investigate how such processes supported microbial metabolism within the persisting shale community, we coupled MAG metabolic profiles with recovered fluid metabolite chemistry. Bacteria and archaea that encode expansive CAZyme profiles are likely capable of degrading polymers such as guar gum and cellulose – some of the most common organic polymers present in frack fluids[59,60]. In the STACK system, we infer that multiple taxonomically distinct fermenters – primarily Thermotoga petrophila (M2-7-6-bin.8), Clostridia SK-Y3 (K-7-4-bin.6), and Fusobacteriales (K-7-4-bin.55) – were responsible for initially degrading the complex carbon polymers added as amendments (**Figures 2.6 and 2.4**). The potential for guar gum degradation was inferred from the presence of alpha-galactosidases that remove galactose side chains and beta-mannosidases that subsequently

cleave the mannose backbone. Likewise, the ability to degrade cellulose was determined from the presence of CAZymes capable of cellulose backbone and oligo cleavage (**Figures 2.6 and 2.4**). Beyond these specific organic polymers, we detected genes encoding extensive collections of CAZymes (**Figure 2.4**) within many putative fermenters, indicating the capability for the degradation of other minor organic polymers introduced in the HF process.

The degradation of polymeric carbon by fermentative community members yielded a range of waste organic acids that likely fueled respiratory metabolisms through intracommunity metabolic exchange. Acetate production was predicted for the majority of MAGs encoding likely fermenters, and concentrations were observed to increase up to 7 mM in STACK-16 & 17 samples (Figure 3). Similarly, high propionate concentrations (up to 600 μM) measured in STACK-16 & 17 samples likely resulted from the activity of dominant Thermotoga and Clostridia microorganisms (**Figure 2.5**). Reflecting its role as a dominant genome in the STACK samples, the Thermotoga MAG encoded genomic potential for degradation of cellulose, guar gum, and xyloglucan, and its relative abundance was predictive (via sparse Partial Least Squares regression analyses; sPLS) of acetate, propionate, and butyrate metabolite concentrations in the fluids, findings consistent with the metabolic role predicted from the genome. Other significant sPLS linkages between genomes and organic acids were identified for MAGs affiliated with the Fusobacteriales, Clostridiales, and Desulfomicrobiaceae (**Figure 2.5**), further supporting our genomic inferences of carbon cycling in this ecosystem.

While fermentative metabolisms are dominant in this system, we also observed the presence of a lower abundance sub-community of respiratory sulfur reducing microorganisms. Freshwater used in the hydraulic fracturing process can promote the dissolution of sulfate minerals from the surrounding rock matrix[61] and thus produced fluids frequently contain sulfate

and thiosulfate. The organic acids that are generated as waste products from fermentative microorganisms likely serve as electron donors to support this respiratory lifestyle (**Figure 2.7**). Specifically, the presence of putative sulfate- and thiosulfate-reducing microorganisms likely drives consumption of organic acids such as acetate and lactate (**Figure 2.5 and 2.7**). Ultimately, we identified the genome-resolved metabolic potential to catalyze the flow of carbon from added complex organic polymers used in the HF process to the consumption of organic acids by inferred sulfate- and thiosulfate-reducing microorganisms. This finding further emphasizes the importance of input materials in sustaining the persisting microbial community for extended periods of time.

2.3.4 Active viral predation influences microbial community heterogeneity.

Viruses were prevalent in STACK samples, with 5,587 viral contigs (>10kb in length) identified across all produced fluid and input samples. The majority of viruses detected in this study were identified from topside input samples, with 748 found to persist in produced fluids recovered from the STACK shale play. The viral populations between wells encompass a majority of shared vMAGs, likely reflecting the previously noted microbial community convergence. However, we also detected subsets of vMAGs unique to each well (**Figure 2.8**) that could be reflective of unique genera, species, or strains that are not shared across wells. Prior to this work, 1,838 vMAGs (>5kb in length), with only 852 >10Kb from 33 samples across 5 HF wells were recovered from the Appalachian Basin[24]. Indeed, only 17 of the viruses recovered from STACK samples were shared with Appalachian Basin vMAGs, and thus our results greatly expand the virome sampling of geographically distinct hydraulically fractured shale ecosystems.

The unique viral populations scale" in 'roportion with the richness of MAGs in each well. Here, STACK-14 hosted the largest number of unique vMAGs and also exhibited the highest microbial host genomic richness. Of the 539 vMAGs that clustered with International Committee on Taxonomy of Viruses (ICTV)-classified reference sequences, all were classified within the viral order *Caudovirales*. Within *Caudovirales*, the majority were in the order *Siphoviridae* (39.5%) followed by *Myoviridae* (34.5%) (**Figure 2.8**). However, the majority of vMAGs identified in these STACK samples could not be assigned to ICTV taxonomic clusters, highlighting the novelty of viruses present in this engineered deep terrestrial ecosystem. Responding to the presence of these viruses, the majority (18 of 24) of the dominant MAGs, including every MAG that achieved 20% or greater relative abundance in a given sample, encoded a CRISPR-Cas viral defense system (**Figure 2.9**). Furthermore, only one MAG, a low relative abundance *Desulfomicrobiaceae* (WD-3-bin.38) that was present at the last sampling time point (~500 days), lacked a CRISPR-Cas system. Through the perfect matching of viral protospacers (i.e., sequences in vMAGs) with spacers in bacterial CRISPR-Cas systems we directly linked viruses to 12 microbial hosts, with the majority of MAGs linked to multiple vMAGs (**Figure 2.9**). The identification of CRISPR-Cas-protospacer matches between viruses and half of the persisting bacterial hosts highlights the extent of virus-host interactions in this subsurface ecosystem and the role these processes likely play in shaping community assembly.

Our findings also provide new insights into viral ecology of this system. We report an instance where the same virus was linked to two distinct *Firmicutes* MAGs*, Peptococcia* and *Clostridia* SK-Y3 (M1-7-4-bin.22 and K-7-4-bin.6, respectively). Identification of the protospacers from both MAGs that were linked to the virus revealed that they were not identical, and matched viral genes for phosphopentomutase and a helix-turn-helix domain protein. We

consider this observation likely the result of incorporation of protospacer sequences from common viral genes (likely from two distinct viruses) into bacterial spacer arrays, rather than multiple infections from the same virus with a broad host range. Viruses generally exhibit high host specificity and infection across multiple different genera is uncommonly reported using similar methods[62,63].

Bacterial interactions with viruses, as inferred from CRISPR-Cas linkages, had variable impacts on the ability of a given MAG to persist within the STACK ecosystem. For example, a *Lachnospirales* MAG (M2-7-5-bin.8) that was linked to seven unique viruses exhibited dramatic decreases in relative abundance across all three wells – a common characteristic of microorganisms under viral predation in many other ecosystems[24,64,65] (**Figure 2.9**). In contrast, the dominant *Thermotoga* MAG was linked to two viruses yet generally did not exhibit relative abundance decreases (**Figure 2.9**) between the timepoints sampled. We note however, that MAGs are composites of many populations of closely related members, and thus the impact to specific strains may be obscured in this approach.

It is likely that many taxa are impacted by viral predation in this ecosystem. Evidence in support of this is the positive correlation between the most abundant MAGs (e.g., *Thermotoga*) and the relative abundance of viruses that are linked to them (**Figure 2.9**). Given the requirement of active bacterial cells for viral replication, these patterns imply that dominant microbial taxa must be continually infected and lysed to support these large pools of free viruses. Additionally, cell lysis can result in mobilization of key metabolites that can subsequently act as substrates for the remaining microbial community[24]. We previously observed such processes occurring at the strain level in samples recovered from Appalachian Basin shales, where infection and associated cell lysis of one *Halanaerobium congolense* cultivated strain yielded niche space for emergence

of another distinct strain[24]. However, these dynamics can be obscured at higher taxonomic (e.g., species, or MAG) levels, resulting in the appearance of stable community composition. Here, we anticipate that similar virus-host interactions are occurring in *Thermotoga*, resulting in an ongoing 'arms race' between multiple *Thermotoga petrophila* strains and associated viruses that supports high relative abundances of both virus and host.

2.3.5 Lower salinity deep shales are characterized by higher taxonomic and metabolic diversity and the dominance of Thermotogae

To date, the majority of genome-resolved metagenomic studies detailing the microbiology of HF systems were performed in eastern US shale formations (i.e., Marcellus & Utica formations). These systems distinguish themselves from the STACK shale play through the presence of highly saline produced waters, generated from the dissolution of salt minerals in the shale rock[14,28,66]. For example, produced water in the Appalachian Basin can reach brine-level salinities (126.74 ± 35.61 mS/cm), whereas salinities in the STACK produced fluids were roughly 5-fold lower (25.06 ± 8.85 mS/cm). Although accurate temperature measurements for hydraulically fractured wells can be difficult to obtain, it is likely that the STACK wells also exhibit higher temperatures compared to their eastern counterparts[29,67].

Due to thermodynamic and physiological constraints, salinity likely exerts a strong influence on the microbial community within the shale fracture network. Consistent with this concept, we measured 4-fold higher Shannon's diversity in these less saline STACK wells, relative to microbial communities in produced fluids from Appalachian Basin wells (**Table 2.2**). However, as generally observed across the majority of time-resolved shale studies, microbial

alpha diversity in STACK samples decreased with time, reflecting the influence of other abiotic and viral constraints on community assembly through the lifetime of the wells (**Figure 2.3**).

Salinity can also constrain the ability of specific metabolisms (and therefore taxonomies) to operate in a given environment[25–27]. For example, heterotrophic sulfate reduction may not be thermodynamically favorable in environments where the cost of osmoregulation is greater than the energy gained from a redox couple. This principle was previously used to explain the absence of canonical sulfate reducing microorganisms from high salinity wells in the Appalachian Basin[5]. In contrast to those results, here we observed a persistent, low relative abundance community of inferred respiratory sulfate- and thiosulfate-reducing microorganisms in the STACK wells that are likely able to tolerate the lower salinity conditions (**Figure 2.6**). Further underscoring the contrasts between these two basins, we also note the lack of genomic potential for the cycling of quaternary amines and methylamines in the lower salinity STACK shale play (Supplementary Discussion). However, despite lower salinities relative to the Appalachian Basin, we still observe the prevalence of osmoprotection strategies in the dominant STACK MAGs, suggesting the importance of this physiological trait in persisting in this ecosystem (**Figure 2.10, Table 2.3, Appendix A**).

Finally, we note the impact of salinity on the distribution of *Thermotoga* species across HF shales. As described earlier, a *Thermotoga petrophila* MAG (M2-7-6-bin.8) was dominant in the majority of STACK produced water samples (**Figure 2.3**). This is in contrast to the majority of samples from the Appalachian Basin where halophilic fermenter *Halanaerobium* strains dominate the microbial communities and likely occupy similar niches in the shale ecosystem[5,13,22,68,69].

Expanding these analyses, we assessed the geographic distribution of *Thermotoga* across a range of fractured shales displaying gradients in salinity and temperature[18,29,67]. Equipment used in the drilling and development of HF wells is re-used across large geographic areas, potentially aiding in the distribution of dominant microorganisms such as *Thermotoga* and *Halanaerobium*. However, as shown here, neither of these taxa dominate in all shale formations. To better understand this pattern, marker gene (i.e., 16S rRNA gene) relative abundance data from this study was paired with results from existing deep shale ecosystem studies from the Utica[5] and Marcellus formations[7,16], Bakken [17,18] and Three Forks formations, and the Denver-Julesburg (DJ) Basin [18]. Our analysis revealed that *Thermotogae* displays a clear biogeographic signal, decreasing dramatically in relative abundance as formation salinity increases to values characteristic of Appalachian Basin or Bakken formations (**Figure 2.11**). These observations suggest that *in situ* salinity may act as a control on *Thermotoga* distribution across the deep terrestrial subsurface. Temperature also likely has an effect on microbial community composition in different shale formations. In contrast to the effects of salinity, elevated temperatures are known to select for thermophilic taxa such as *Thermotoga*[70,71]. As such, the presence of higher *in situ* temperatures in the STACK formation (100-120°C)[29] compared to the Appalchian Basin (50-100°C)[67], likely promotes *Thermotogae* dominance in this system. We speculate that in more saline shale ecosystems, *Thermotoga* may be unable to compete with *Halanaerobium*, while in the presence of elevated temperatures and lower salinity (i.e., STACK shale play, DJ Basin), *Thermotoga* may out-compete *Halanaerobium* with lower temperature growth thresholds.

Despite the differences in microbial community composition between the STACK shale play and the Appalachian Basin, the dominance of either *Halanaerobium* or *Thermotoga* highlights the central conserved role that fermenters play both in these ecosystems. These

halophilic or thermophilic taxa may be thought of as 'microbial weeds' that encode specific traits, allowing them to maximize the conditions and available resources in the aftermath of HF and out compete other microbial community members[72]. We infer that the ability to degrade complex carbon polymers used in the HF process is a key trait for microorganisms to persist in fractured shale ecosystems. Although *Thermotoga* contained CAZymes for degradation of common topside additives (e.g., Guar Gum), this MAG contained fewer CAZymes than many of the other inferred fermenters in the STACK system. These observations suggest that in the relatively stable chemical environment of the deep subsurface, a more constrained genomic repertoire may be optimal for persisting over extended periods of time[73], in contrast to other 'opportunitroph' microorganisms with broader metabolic and physiological potential[74].

2.4 Conclusions

This study used a paired metagenomic and metabolomic approach to expand the genomic memberships and metabolisms known to occur within fractured shale ecosystems. The closed nature of these environments and the novelty of topside samples obtained in this study allowed us to trace key input chemistry to the pool of resulting microbial metabolites and develop a framework for metabolic exchanges between fermenters and respiratory organisms that sustain the persisting microbial community. Additionally, this study is the first in performing genome-resolved source tracking to determine the potential topside sources of key taxa that inoculate hydraulically fractured ecosystems, revealing that source waters, fracture fluids, and drill muds are likely areas for future microbial control. We also contrasted the taxonomic and functional profiles in the STACK shale play with well-characterized microbial communities from shales within the Appalachian Basin. Notably, we observed relatively high microbial alpha diversity in

the STACK shale play, as well as the absence of *Halanaerobium*, and greatly expanded the fractured shale virome. Finally, a meta-analysis of microbial community data revealed the impact of physicochemical conditions (i.e., temperature, salinity) on the ability of specific 'opportunitrophic' taxa to dominate shale ecosystems, suggesting that *Thermotoga* could play a dominant role in other low salinity systems. Together, these insights offer a better understanding of the effects of deep biosphere physiochemical conditions on colonization, persistence, and microbial community dynamics in a newly formed subsurface ecosystem.

2.5 Materials & Methods

2.5.*1 Input, flowback, and produced fluid sampling*

Produced and flowback fluid samples (*n = 18*) were collected from three hydraulically fractured shale wells in the STACK shale play, within the Anadarko Basin, which is an important reservoir of oil and gas (Oklahoma, USA)[32] (**Figure 2.1**). Target formations within the STACK shale play include the Woodford and Meramec formations, which are located at approximately 2,440-3350 meters depth in the subsurface where temperatures likely range between 100-120°C. Two wells (STACK-16 & 17) were adjacent to each other on the same well pad and were hydraulically fractured by the same company and consequently received nearly identical chemical additives. In contrast, while the STACK-14 well was present in the same shale play, it was located approximately 10.5 km from STACK-16 & 17 and hydraulically fractured by a different company, resulting in a different suite of chemical additives (**Appendix B**). All three HF wells of interested (STACK 14, 16, & 17) were hydraulically fractured in August 2017. Temporal sampling was conducted for each of the three wells over approximately a 400 day timeseries as follows: STACK-14 (*n=7*) from days 80-571 post-hydraulic fracture, STACK-16

(*n=5*) from days 141-514 post-hydraulic fracture, and STACK-17 (*n=6*) from days 105-514 post-hydraulic fracture. In addition to the sampling of produced and flowback fluids, we obtained 10 unique types of top side samples (*n=13*) (e.g., source waters, biocide-treated waters, frack fluids) used during the development and HF process for both sets of wells (**Figure 2.1**). Drill mud samples were obtained at the time of the HF of STACK 16 &17 from an adjacent well pad. A limited number of produced fluid samples were also recovered from two other nearby STACK wells, STACK-12 (*n=3*) and STACK-13 (*n=3*). However, due to the small number of produced fluid samples, lack of associated input samples, and limited general information about these wells, STACK-12 and STACK-13 were only used for the recovery of MAGs and were otherwise excluded from our analyses. Input materials for each well were sampled at their sources throughout the HF process, while HF produced and flowback fluids were collected from the gas-water separator tanks associated with each well in 1L sterile Nalgene bottles with no headspace. All samples were shipped overnight on ice and filtered through a 0.22 µm filter upon arrival. The filter was stored at -80°C until DNA extraction.

2.5.2 Chemical and metabolite analysis

Conductivity was measured on raw, unfiltered fluids using a Myron L 6PIIFCE meter. Filtrate of all 0.22-µm filtered fluid samples (input and produced/flowback fluids) were sent to the Pacific Northwest National Laboratory for Nuclear Magnetic Resonance (NMR) spectroscopy metabolite analysis. Samples were diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 as an internal standard. All samples were analyzed in 3 mm NMR tubes at a regulated temperature of 298 K. Chemical shifts were referenced to the ^{1}H or ^{13}C methyl signal in DSS-d_6 at 0 ppm. The 90° ^{1}H pulse was calibrated prior to measurement

of each sample. The one-dimensional ^{1}H spectra were acquired using the Varian tnnoesy pulse sequence with a spectral width of 12 ppm and 512 transients. A pre-saturation pulse was applied to water for 1.5 s prior to the start of the sequence, the NOE mixing time was 100 ms and the acquisition time was 4 s. Time domain free induction decays (57472 total points) were zero-filled to 131072 points prior to Fourier transform. Candidate metabolites present in each of the complex fluid mixtures were determined via matching of the chemical shift, J-coupling, and intensity information of experimental NMR signals against the NMR signals of standard metabolites in the Chenomx, against compound signals in the Chenomx, Human Metabolome Database (HMDB), and custom in-house databases. Additionally, 2D spectra (including ^{1}H–^{13}C heteronuclear single-quantum correlation spectroscopy, ^{1}H-^{1}H total correlation spectroscopy) were acquired on most of the fluid samples, aiding in the 1D ^{1}H assignments of several metabolites. Signal to noise ratios (SNR) were measured using MestReNova 14 with the limit of quantification equal to a SNR of 10 and the limit of detection equal to a SNR of 3. TOCSY and HSQC spectra were processed and ^{1}H and ^{13}C chemical shifts identified using MestReNova[14]. Complete metabolite and conductivity data for samples paired with metagenomics is provided in **Appendix B**.

2.5.3 DNA extraction and metagenomic sequencing

Samples of 500-800mL fluids from inputs and produced fluids taken at separator tanks were filtered through 0.22-µm pore size polyethersulfone filters (Millipore, Fisher Scientific). Input frack fluids were too viscous for our filtration protocol and thus 50-100 mL of the solution was spun down in a centrifuge at 8000xg for 1 hour to pellet solids. Supernatant was removed and DNA was extracted from the remaining pellets. Drill muds also could not be filtered and

therefore DNA was extracted directly from 5 mL of these materials, without pelleting or filtration. Total nucleic acids were extracted from the filters, frack fluids, and drill muds using *Quick*-DNA Fecal/Soil Microbe Microprep Kit (Zymo). All rounds of DNA extractions were performed with corresponding extraction blanks to ensure no contamination occurred during the laboratory extraction process. All blanks returned no detectable nucleic acids using the maximum amount of blank sample (20μL) according to the Qubit dsDNA High Sensitivity assay kit (ThermoFisher Scientific). All inputs, with the exception of one frack fluid used in STACK-16 & 17 (MC-6-SW), and most produced fluid samples were sent to the Department of Energy Joint Genome Institute (JGI) for library preparation and sequencing. One frack fluid used in STACK-16 & 17 (MC-6-SW) was sequenced at Ohio State University's Comprehensive Cancer Center Genomics Shared Resource using the Illumina Nextera XT Library System according to manufacturer's instructions. The remaining produced fluid samples were prepared and sequenced at the Genomics and Microarray Core at the University of Colorado, Denver's Genomics Shared Resource. For samples sequenced at JGI, an Illumina library was constructed and sequenced 2x151 using the Illumina HiSeq-2500 1TB platform and paired-end reads were collected. Samples sequenced at the University of Colorado, Anschutz Medical Campus were prepared using the Illumina Nextera XT Library System according to manufacturer's instructions for 2x151bp libraries. Libraries were sequenced using the Illumina NovaSeq platform and paired-end reads were collected. Information about sequencing for each sample is listed in **Appendix B**. Raw sequences were deposited to NCBI under BioProject PRJNA30832.

2.5.4 16S rRNA gene sequencing and analysis

Nucleic acids for all samples were also sent to Argonne National Laboratory for 16S rRNA gene sequencing (**Appendix B**). Due to challenges in recovering DNA from many HF inputs, 16S rRNA gene sequencing and analysis includes technical and biological replicates for many topside samples that did not undergo metagenomic sequencing. Further information on sample treatment can be found in **Appendix B**. Sequencing was performed with the Illumina MiSeq platform, using the Earth Microbiome Project barcoded primer set (forward primer, 515F, 5'-GTGYCAGCMGCCGCGGTAA-3'; reverse primer, 806R, 5'-GGACTACHVGGGTWTCTAAT-3') to amplify the 251bp hyper-variable V4 region. 16S rRNA gene sequences were obtained via Argonne's standard procedure, with the exception of performing 30 PCR amplification cycles. Paired-end reads were processed with QIIME2 (v 2019.7) EMP protocol, by first demultiplexing via exact-match of barcodes, trimmed to 250bp and denoised with DADA2[33], and then taxonomically classified with SILVA (release 132). All 16S rRNA gene sequencing reads were submitted to NCBI under BioProject PRJNA30832 and individual accession numbers are listed in **Appendix B**.

2.5.5 Metagenomic assembly, binning, and analysis

Total sequenced DNA from each sample was first trimmed from 5' to 3' ends with Sickle (https://github.com/najoshi/sickle) and individually assembled using IDBA-UD with default parameters[34]. Assembly information for each sample is provided in **Appendix B**. Scaffold coverage was determined by read recruitment back to assemblies, via BowTie2[35]. Only scaffolds >5kb from metagenomic assemblies were binned with MetaBAT2 to recover metagenome assembled genomes (MAGs)[36]. Produced fluid samples from STACK-12 ($n=3$) and STACK-13

(*n=3*) were included to build a comprehensive database of STACK-curated MAGs, but were not included for other metagenomic analysis. CheckM (v.1.1.2) lineage workflow ('lineage_wf') followed by the 'qa' command was used to assess completion and contamination for each metagenomic bin[37], and a total of 646 medium (>50% completion, <10% contamination) and high (>90% completion, <5% contamination) quality bins were recovered from all input and produced fluid samples (*n = 31*), following the standard metrics for MAGs proposed by Bowers *et al*.[38]. The curated STACK database of 316 unique MAGs was determined by dRep v2.2.3[39] using default parameters.

All MAGs were taxonomically classified using GTDB-Tk v1.0.2[40]. Metagenomic assemblies were annotated via DRAM v1.0.5 using default parameters[41]. The recruitment of metagenomic reads was used to infer MAG relative abundances across all time points. To determine relative abundances of MAGs and thus temporal dynamics in all three wells metagenomic reads were first randomly sampled up to 13Gbp using bbtools to account for varying sequencing depths[42], and then multi-mapped to all 316 unique STACK genomes via BowTie2, with minimum scaffold coverage of 75% and depth of 1 required for read recruitment (https://github.com/TheWrightonLab/metagenome_analyses). For a MAG to be considered present in any given sample, each MAG needed to have >90% of its scaffolds with >1 coverage. Relative abundances for each MAG were calculated as their coverage proportion from the sum of the whole coverage of all bins for each set of metagenomic reads. The 24 dominant and persisting (>5% relative abundance in any given STACK-14, STACK-16 or STACK-17 sample) medium and high quality, dereplicated MAGs were deposited at NCBI within BioProject PRJNA30832.

2.5.6 Viral recovery and analysis

Viral MAGs (vMAGs) were identified in metagenomic assemblies using VirSorter[43] within the CyVerse discovery environment. VirSorter was run with default parameters using the 'virome' database and viral contigs with category 1 or 2 (free) and 4 or 5 (integrated) were retained. Viral genomic contigs (≥10kb) were clustered into viral populations (genus level) using the 'ClusterGenomes' (v 1.1.3) app in CyVerse using the parameters 95% average nucleotide identity and 90% alignment fraction of the smallest contig. To calculate the viral relative abundance of viral contigs, BBMap[42] multi-mapped the metagenomic reads to unclustered viral contigs with minimum 90% identity. Next, CoverM (v 0.4.0) calculated the coverage of viral contigs, requiring a minimal scaffold coverage of 75% (https://github.com/wwood/CoverM). To taxonomically classify shale-derived viral contigs in the context of known viral sequence taxonomy we used the database of the International Committee on Taxonomy of Viruses (ICTV), a network-based protein classification was performed[44,45]. Predicted proteins from shale-derived viral contigs were clustered with predicted viral proteins contigs within the NCBI Bacterial and Archaeal Viral RefSeq database (v85) with a required *E* value of 1×10^{-3} and processed using vContact2 (v 0.9.8)[46]. Taxonomy of shale-derived viral contigs was predicted for sequences that co-clustered with reference viral sequences of known taxonomy. If viral clusters exclusively contained shale-derived viral sequences from this study, the viral cluster was termed previously undescribed. To match viral contigs to microbial hosts, CRISPR-Cas arrays were first identified in each bacterial or archaeal genome using the CRISPR Recognition Tool plugin[47] in Geneious (v. 2020.0.5). Then, the identified protospacers from the hosts' CRISPR-Cas array were queried against all viral contigs using BLAST(n) to identify sequences that perfectly matched and make strong linkages between host and viral populations.

2.5.7 Metabolic profiling of STACK MAGs

To assess metabolic potential, MAGs were annotated via DRAM v1.0.5 using default parameters[41]. Results from DRAM annotations were leveraged to make inferences about MAG metabolic potential including fermentative or respiratory lifestyles (Table **Appendix B**). MAGs encoding inferred fermentative microorganisms were identified by a lack of a complete electron transport chain and wide repertoire of carbohydrate active enzymes (CAZYmes). Conversely, MAGs encoding inferred sulfur respiring microorganisms were characterized by the presence of reductive dsrAB (sulfate) and/or phsA and rhodanese genes (thiosulfate), fewer CAZymes and more complete electron transport chains. Putative methylamine related genes (such as cutC, grdI, and mttB) required manual confirmation. Choline trimethylamine-lyase (cutC) genes were confirmed via alignment of sequences with those provided in Craciun et al., 2014, and verification of active sites[48]. The only glycine reductase gene (grdE), was confirmed by aligning and constructing a RaxML tree with sequences from Daly et al., 2016 according to their methods[5]. Trimethylamine methyltransferase (mttB) were confirmed by aligning with known mttB sequences[5]. Lastly, genes related to osmoprotection in all 24 dominant MAGs were identified from a manually curated DRAM distillate sheet using DRAM v.1.2.0 (**Table 2.3**).

2.5.8 Statistical & microbial community analyses

To conduct genome-resolved source tracking, the read-recruitment of input sample metagenomes to MAGs was analyzed. For a MAG to be considered present in an input sample, the same read recruitment requirements were held (minimum of 75% scaffold coverage, depth of 1, with 90% of scaffold with >0 coverage). However, unlike calculations to determine relative

abundances through time of the produced fluids, we utilized the full sequencing depth (not rarified) of input samples to increase the likelihood of detecting a key MAG in these samples. For the purpose of this study, if the MAG made the minimal requirements to be considered present, we deemed that MAG 'detected' in the input sample, but due to the complexity and variability of the input samples, we do not apply relative abundance calculations for our genome-resolved source tracking analysis results. All samples listed as 'inputs' in **Appendix B** were included in the source tracking analysis.

Microbial community diversity statistics were analyzed in R (v 3.6.2) using Vegan (v 2.5-6). Shannon's diversity was calculated using relative abundances data derived from the 13Gbp rarified read recruitment to MAGs. Beta diversity was calculated by a nonmetric multidimensional scaling (NMDS) on the resulting feature table from 16S rRNA gene sequencing analysis using Bray–Curtis dissimilarity indices. Beta-diversity was calculated with 16S rRNA gene data, as opposed to metagenomic data, due to the higher number of samples achieved by sequencing biological and technical replicates, and thus stronger inferences could be made. Multiple Response Permutation Procedure (MRPP) and Analysis of Group Similarities (ANOSIM) were used to determine statistically significant differences between groups. Hierarchical clustering of the 24 dominant MAGs was performed in R, using the package 'pvclust' (v 2.2-0) with 100 bootstraps. To determine significant differences between input chemistry between the two sets of wells, linear discriminant analysis effect size (LefSe)[49] was performed on NMR data from frack fluid samples for each set of wells. Since STACK-16 and STACK-17 were hydraulically fractured at the same time and received the same inputs, they were assumed to be one for this analysis. Finally, sparse PLS (sPLS) was used to investigate the relationship between the 24 dominant, persisting MAGs and metabolites in the STACK shale

play[50,51]. A MAG's predictive ability for a specific metabolite was based off of its respective VIP score, of which only scores >2 were considered[52].

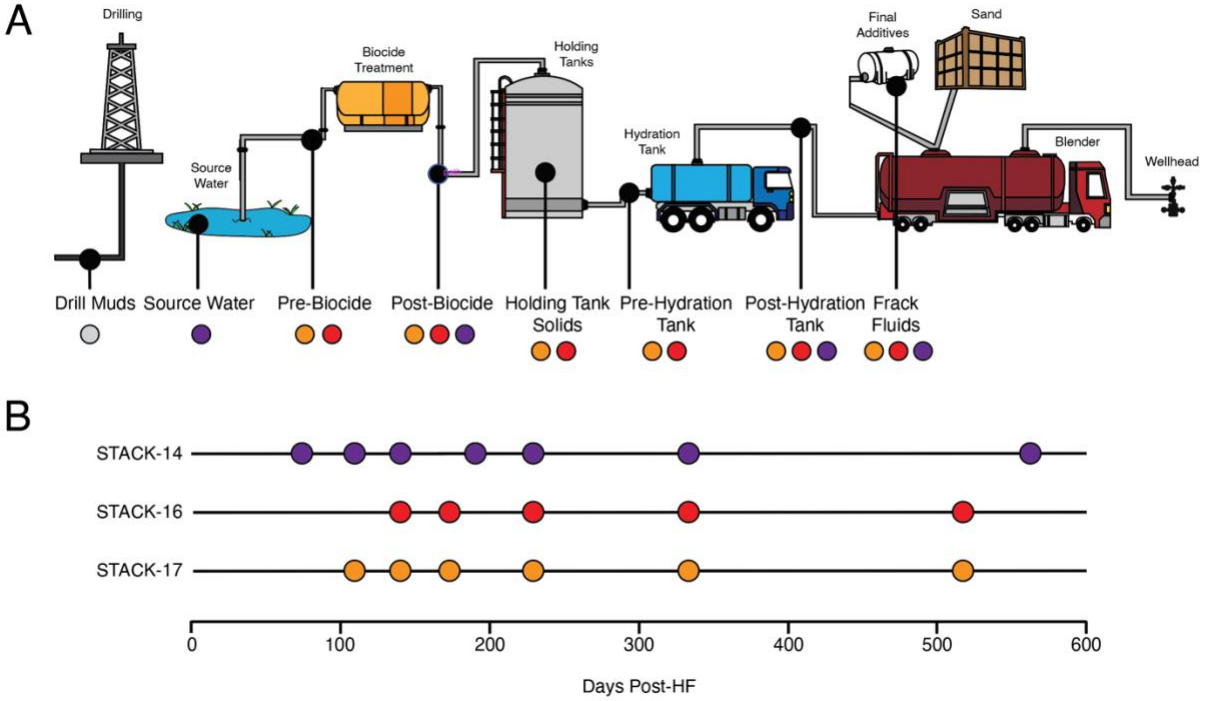

Figure 2.1. Sampling design for input and produced fluid samples with metagenomic sequencing from the STACK shale play. Input samples obtained during the development of the well (A) are color coded to match the produced fluid timeseries for each well (B) in which they are associated with. Drill muds were collected from a nearby, drilling operation and are thus not colored to match the three STACK wells.

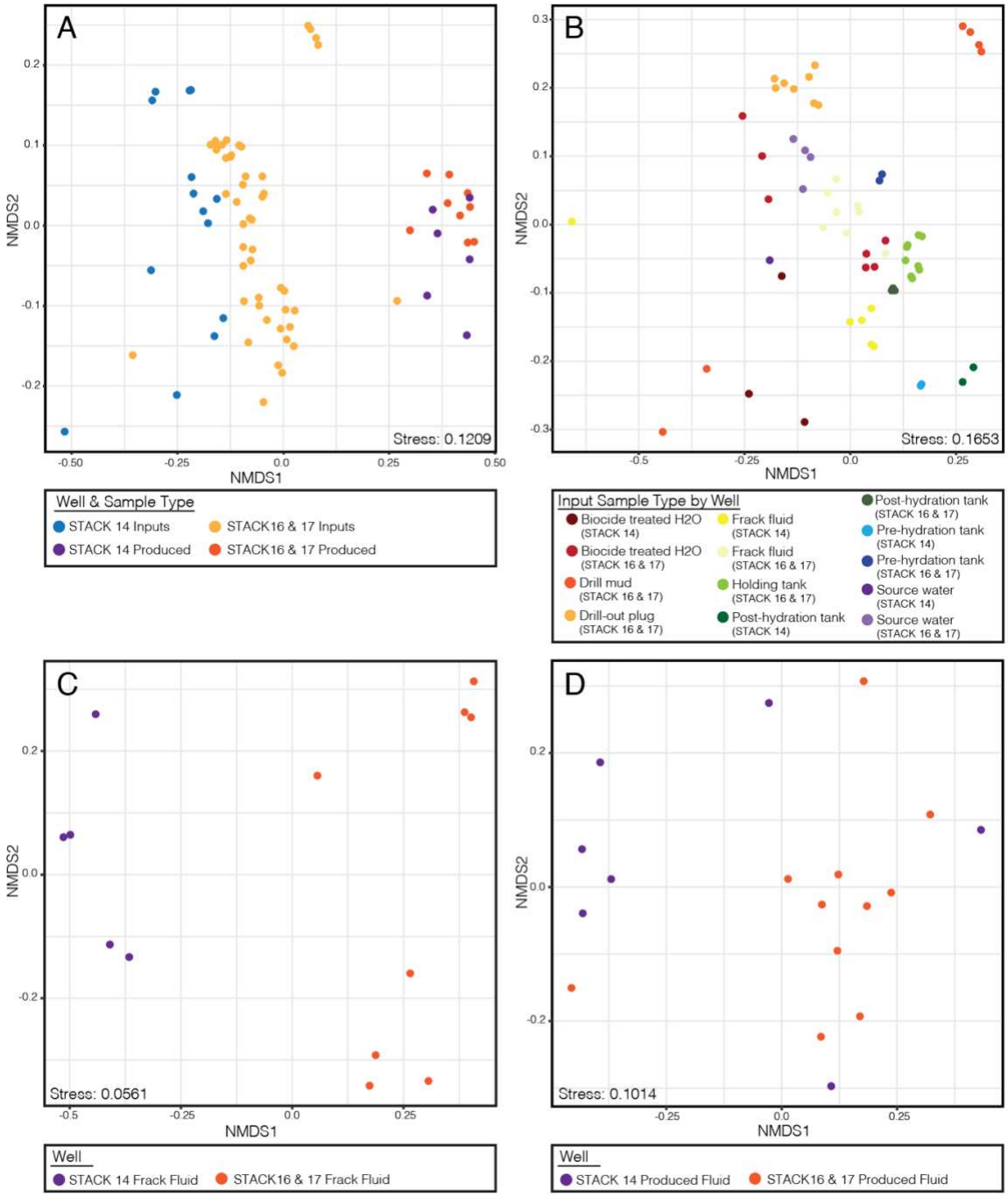

Figure 2.2. Non-metric multidimensional scaling ordination of 16S rRNA amplicon data. **(A)** All inputs and produced fluids for both STACK-14 and STACK-16 & 17, showing a distinct difference between inputs and flowback fluids (MRPP; A: 0.12, p value: 0.001). (**B)** Only input samples for both sets of wells, colored by type of additive used in the development of the HF well (MRPP; A: 0.41, p value: 0.001). **(C)** The resulting final fluids used in HF, frack fluids, are distinctly different between STACK-14 and STACK-16 & 17 (MRPP; A: 0.21, p value: 0.002). **(D)** Microbial communities of flowback and produced fluids for STACK-14 and STACK-16 & 17 not distinct from one another (MRPP; A: 0.05, p value: 0.029).

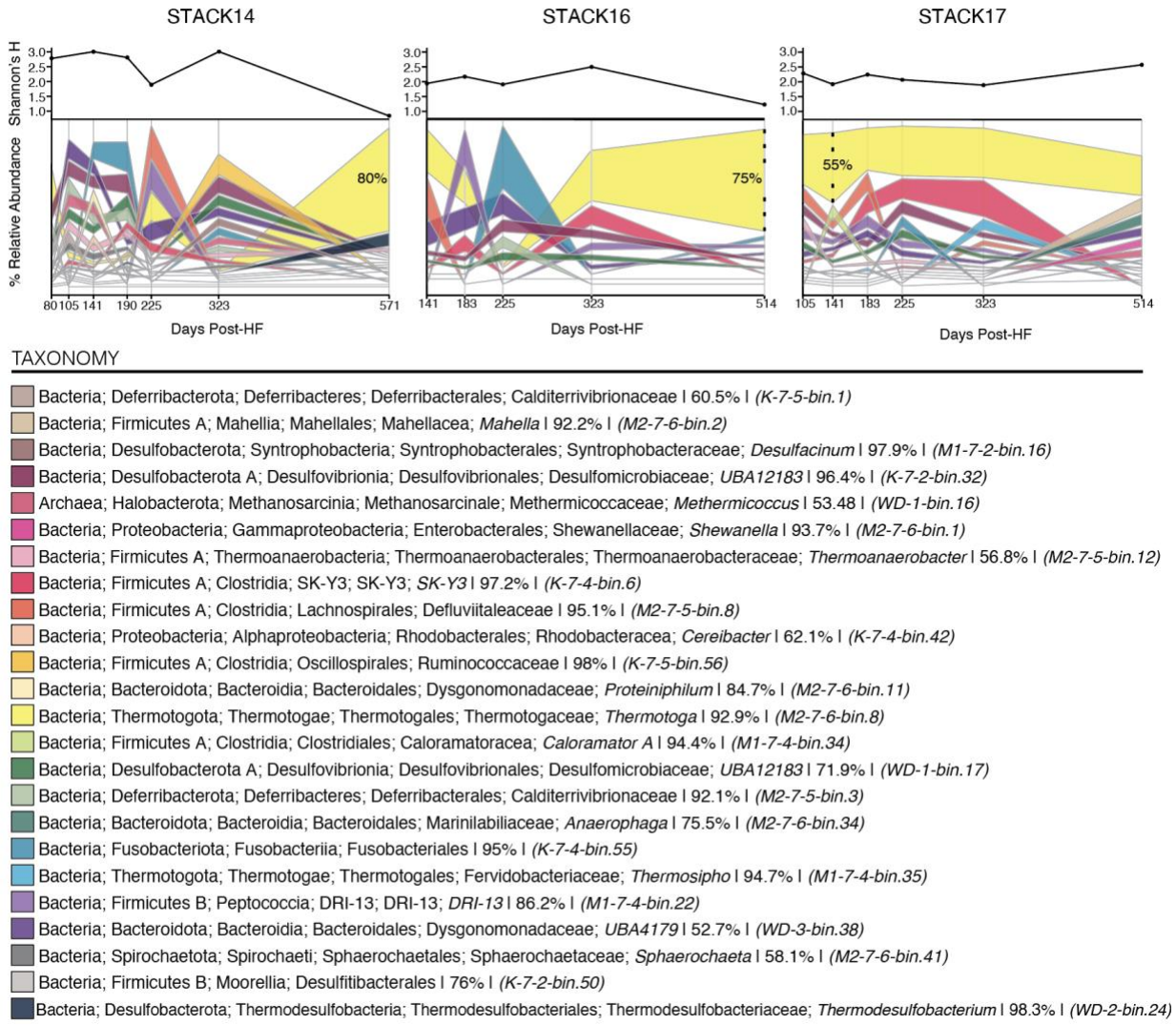

Figure 2.3. Temporal dynamics of the 24 MAGs representing the dominant and persisting taxa (> 5% relative abundance at in at least one sample) in the three distinct STACK shale play wells (STACK-14, STACK-16, STACK-17). Relative abundances were calculated from the metagenomic read recruitment to MAGs as described in the methods. The relative abundance of each MAG is indicated by the width of its respective band in the alluvial plot at each timepoint, with the most abundant MAG on top and least abundant on the bottom and colored by respective taxonomy. Completeness estimates for each MAGs are listed following MAG taxonomy, and unique identifiers for each MAG are listed in parentheses. Trends in alpha diversity through time are shown above each plot for each well.

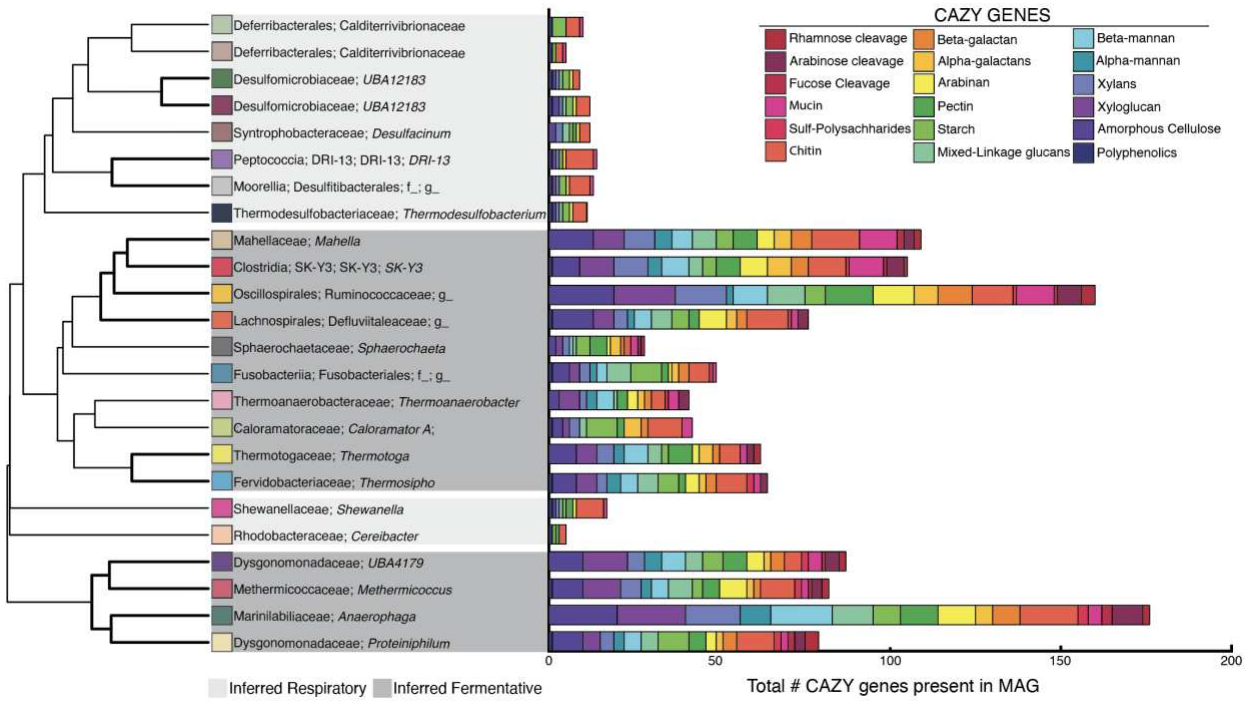

Figure 2.4. Hierarchical clustering of 24 dominant STACK MAGs based off of summarized annotated metabolisms, matched with each's respective # of CAZY genes from DRAM. Bolded black dendrogram branches represent 95% confidence interval groupings. Inferred-respiratory and inferred-fermentative MAGs were determined by completeness of their electron transport chains and CAZY profile.

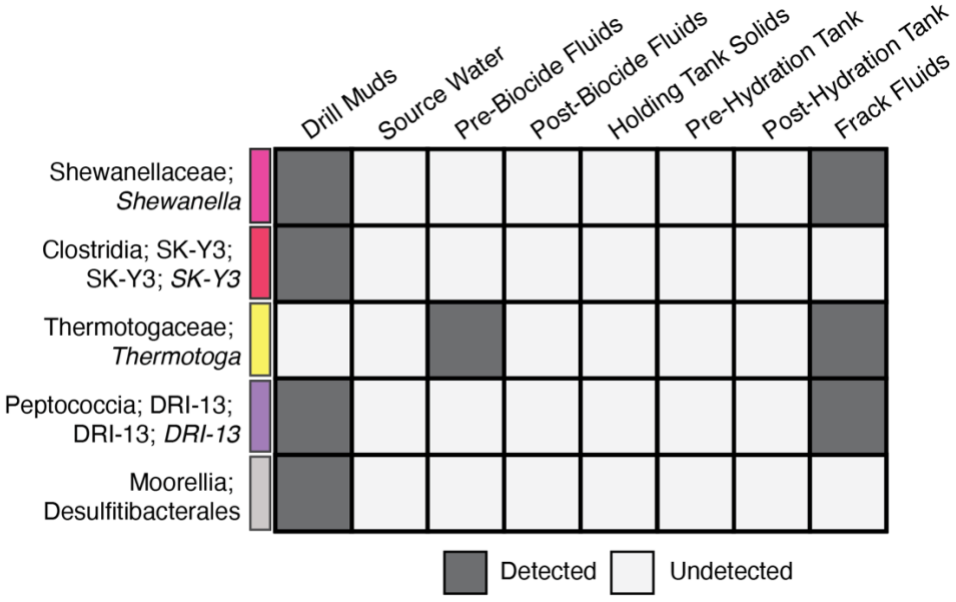

Figure 2.5. Genome-resolved source tracking of 24 dominant STACK MAGs revealed potential topside sources of five key MAGs.

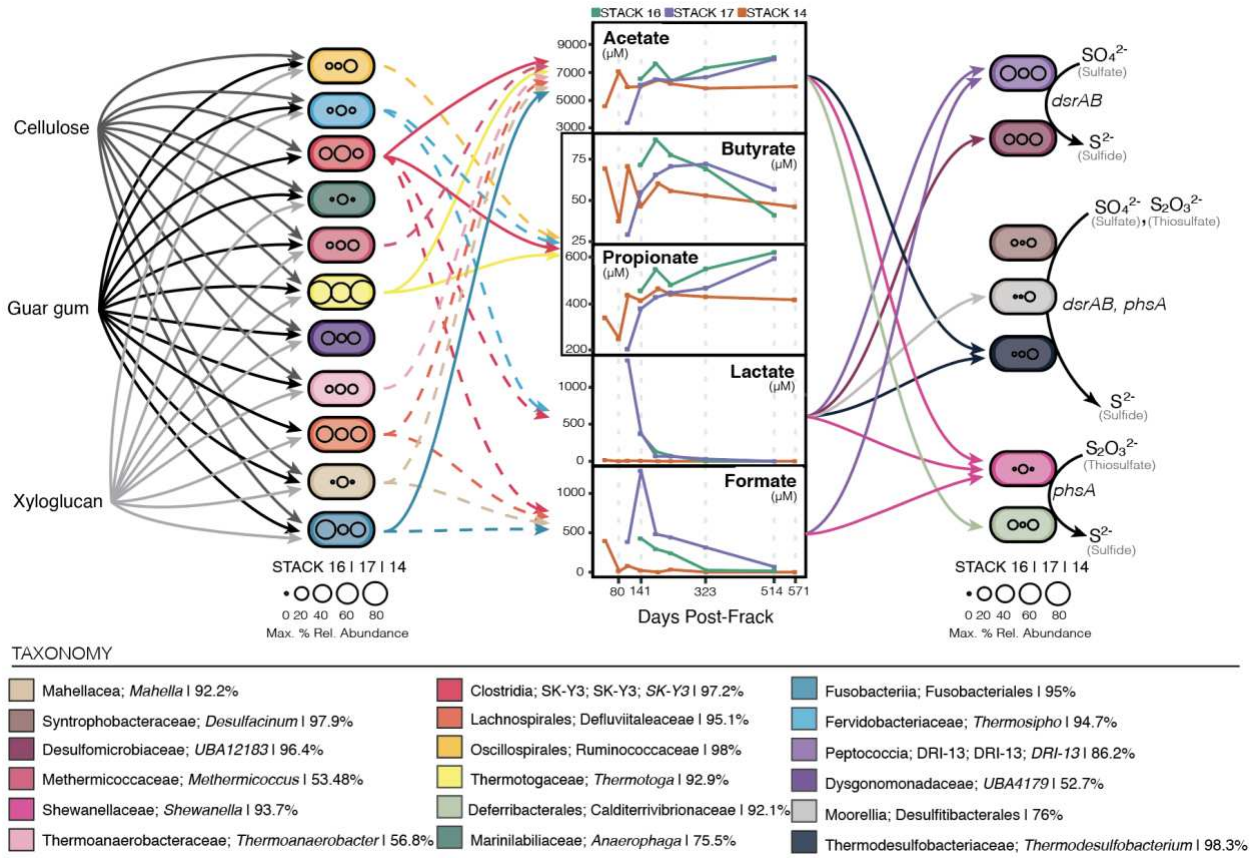

Figure 2.6. Carbon flow in the STACK shale play. From left to right, complex polymers may be degraded by inferred fermentative microorganisms and converted to organic acids, which could be utilized by sulfate- and thiosulfate-reducing microorganisms. Color of each MAG oval corresponds to taxonomic classification, and the size of each circle within each MAG indicates max. % relative abundance for the STACK 16, 17 & 14 wells, respectively. Completeness estimates for each MAG are listed after the MAG taxonomy in the key. All graphs depicting organic acid concentrations are in μM measurements. Solid lines between inferred fermenters (far left) and organic acids indicate genomic potential and statistical prediction to that metabolite via sPLS (VIP>2), while dashed lines only indicate genomic potential. All other solid lines between MAGs and complex polymers, and between organic acids and sulfate reducers, indicate genomic potential for degradation or uptake, respectively.

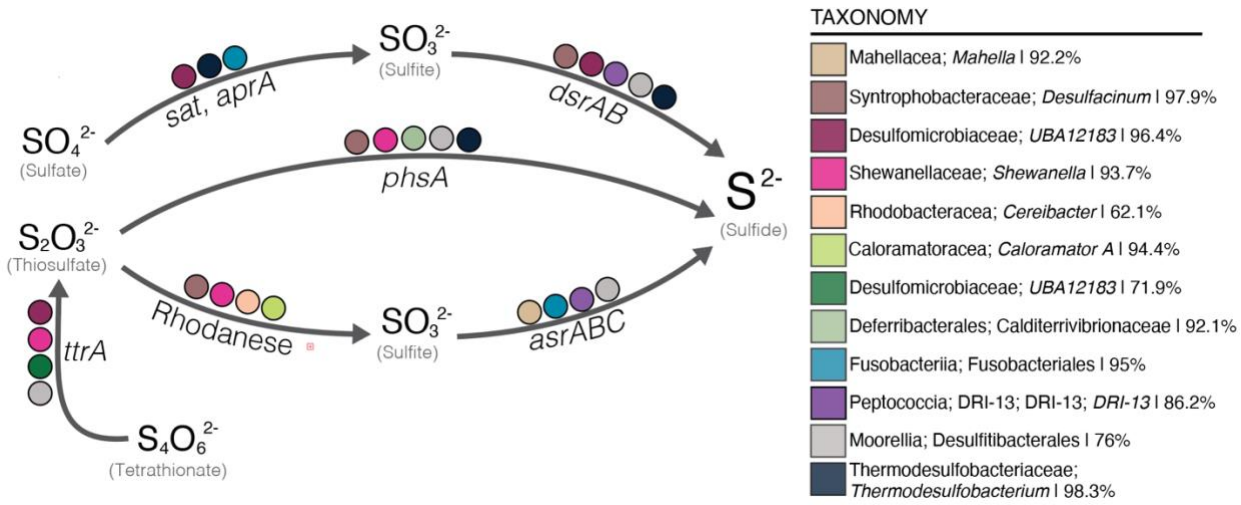

Figure 2.7. Key MAGs encoding taxa inferred to be involved in sulfide generation within the STACK shale play ecosystem. Colored circles indicated the respective MAG contained genomic evidence of the gene in the specified pathway to transform tetrathionate/thiosulfate/sulfate to sulfide. Completeness estimates for each MAG are provided after taxonomy in the key.

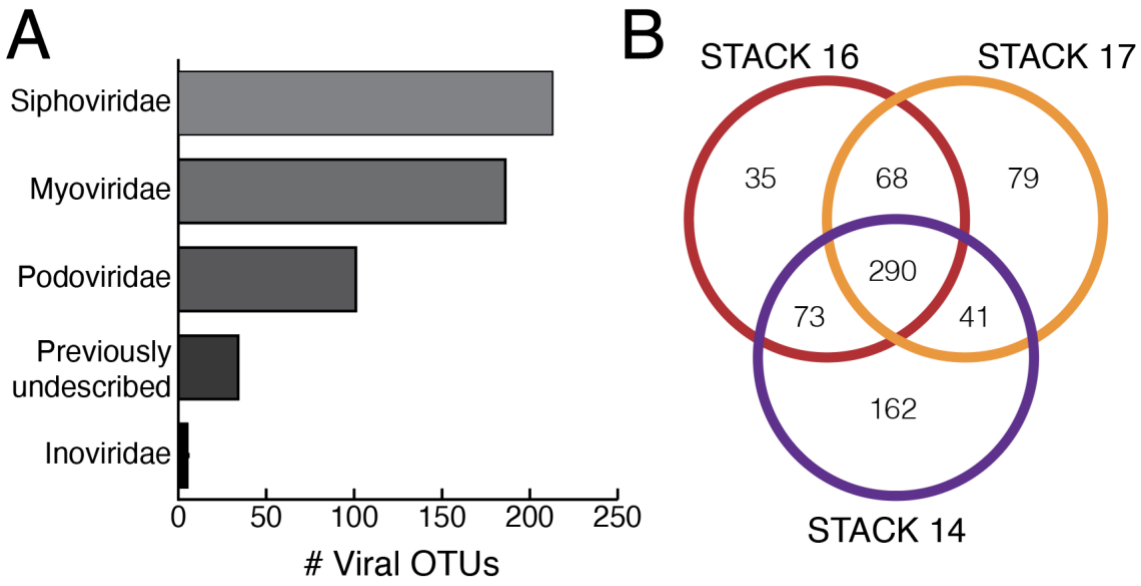

Figure 2.8. **(A)** Family-level viral taxonomy of clustered viral OTUs based on protein clustering and homology. Viral OTUs are considered previously undescribed if the cluster was solely composed of STACK shale-derived viral contigs. **(B)** Venn diagram depicting the number of shared and unique vOTUs per STACK well.

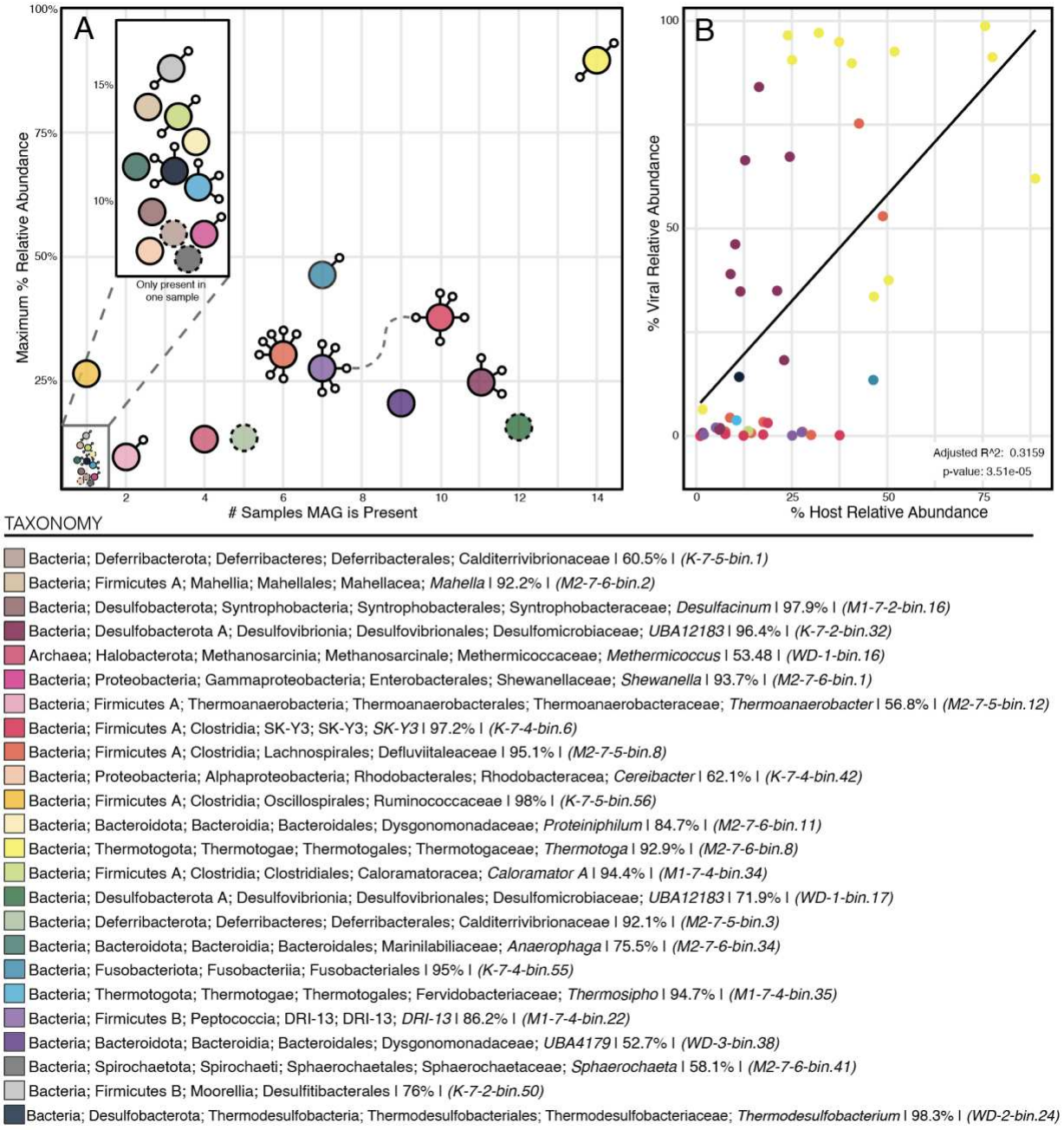

Figure 2.9. Viral-host dynamics in the STACK shale play. **(A)** Visual representation of each of the 24 STACK MAGs 'relevance' and viral connections. Relevance is evaluated by the number of samples where a MAG is present, and the maximum relative abundance that each MAG reaches (considering any given sample). Each MAG is depicted as a colored circle, with a solid line indicating the presence of CRISPR-Cas viral defense system and dashed the absence of one. Small, connected circles represent the viral linkages, and the dashed grey line connecting virus-to-virus indicates an identical spacer sequence (but likely not an identical virus). **(B)** Evaluation of viral and host dynamics where linkages could be made. Relative abundances of hosts and the summed relative abundance of their linked viruses are plotted for each timepoint that the host is present, revealing that the most abundant viruses are associated with the most abundant microbial hosts.

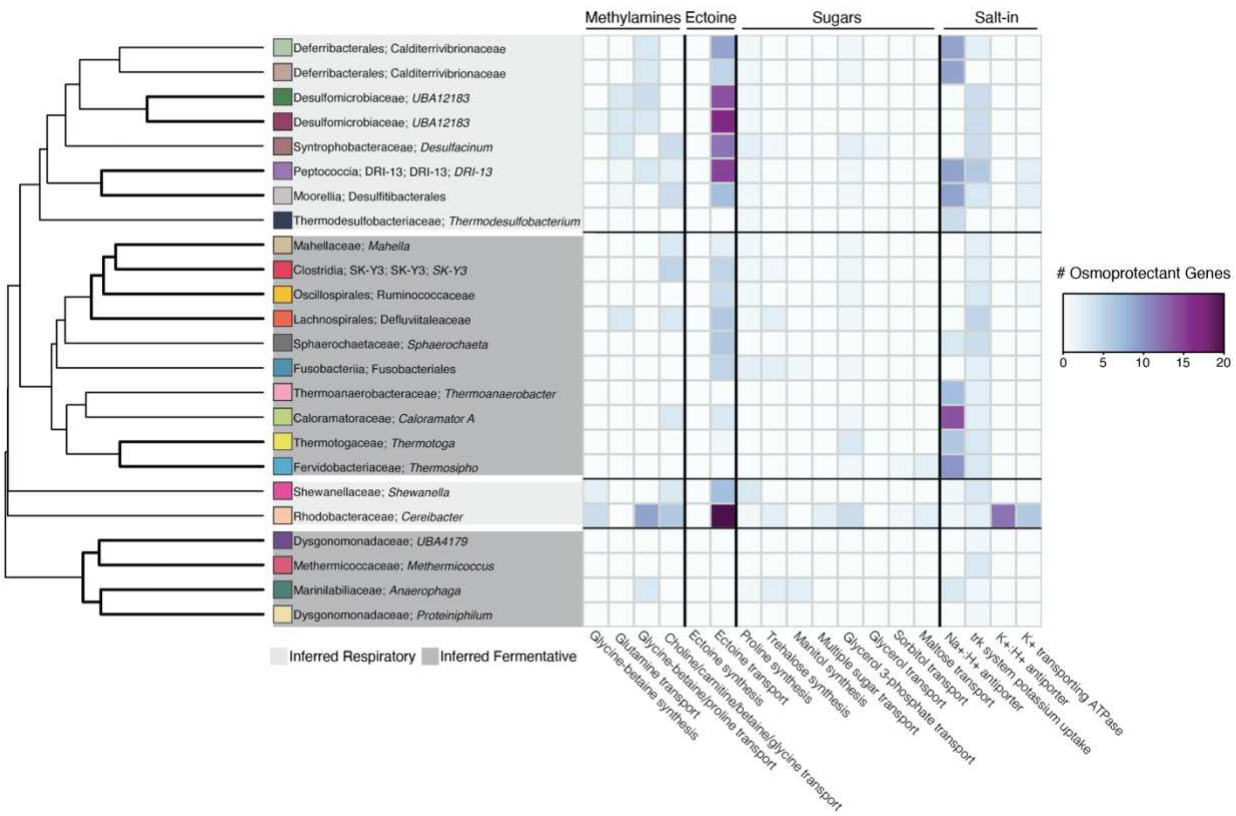

Figure 2.10. Heatmap of osmoprotectant strategies of the 24 dominant MAGs in the STACK formation. MAGs are hierarchical clustered based on DRAM summary of their full annotated metabolism, and bold lines indicate a cluster with a 95% confidence interval. Osmoprotection genes were identified by DRAM annotations and manually categorized (Table S7).

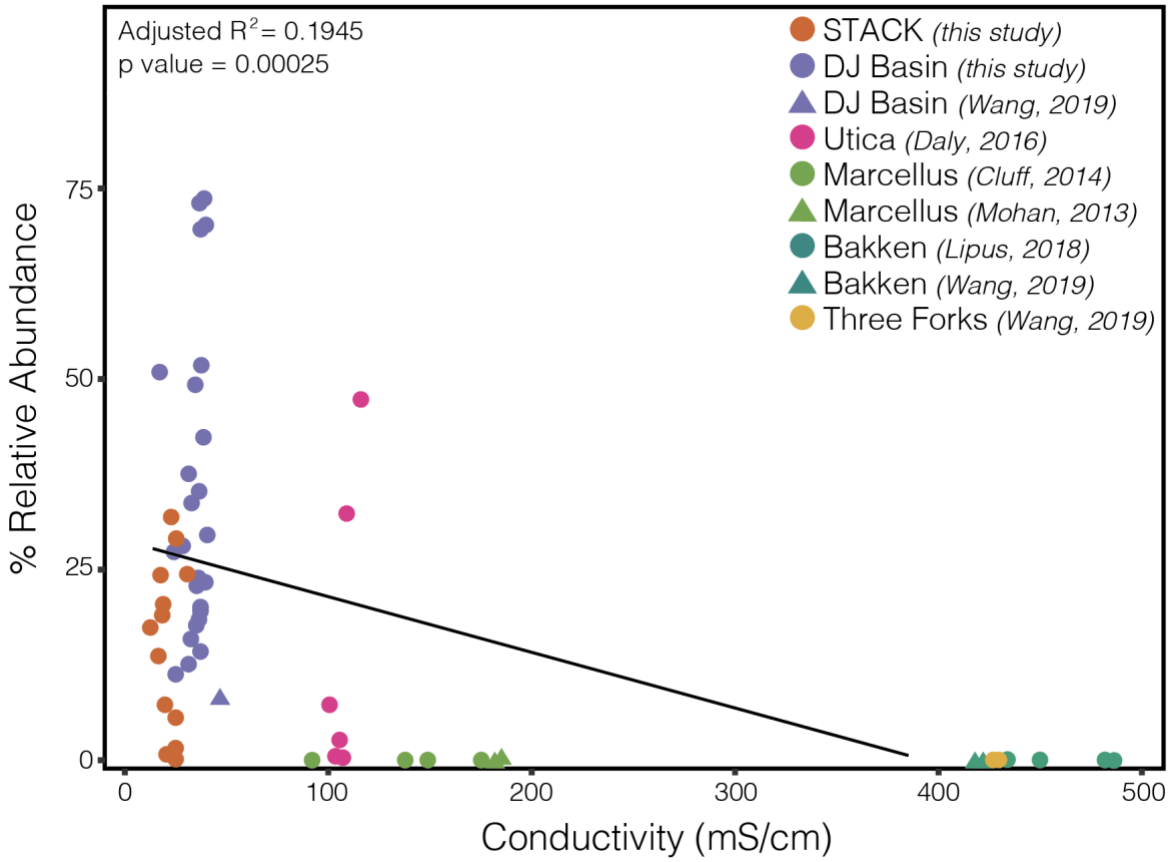

Figure 2.11. 16S rRNA gene relative abundances of *Thermotogae* across US shale formations and their respective salinities, as reported in this study (STACK, DJ Basin) and previously by others (Utica, Marcellus, Bakken, Three Forks).

Chapter 2 Tables

Table 2.1. Linear discriminant analysis effect size (LefSe) analysis of metabolite profiles for input frack fluids for STACK-14 and STACK-16 & 17. Only metabolites that were significantly discriminate of one of the wells are shown.

Feature/Compound	Well for which feature is discriminant	LDA effect size score	p value
Choline	STACK-14	5.175132165	0.049534613
Isopropanol	STACK-14	4.695385691	0.049534613
Acetate	STACK-16 & 17	4.444529653	0.049534613
Glutarate	STACK-16 & 17	4.108332216	0.049534613

Table 2.2. Alpha diversity measurements of both STACK and Appalachian Basin microbial communities (1).

	Shannon's diversity H'	standard deviation
STACK	2.22	0.59
Appalachian Basin	0.61	0.49

Table 2.3. Osmoprotection genes and categorization for **Figure 2.9**.

Category	Figure S5 heatmap gene annotation	Gene name
Salt-in	K+transporting ATPase	*kdpABCDEF*
Salt-in	K+:H+ antiporter	*phaABCDEFG*
Salt-in	trk system potassium uptake	*trkAH, ktrAB*
Salt-in	Na+:H+ antiporter	*mnhABCDEFG* *nhaC*
Sugars	maltose transport	*malFG*
Sugars	sorbitol transport	*smoE, mtlE, srlA*
Sugars	glycerol transporter	*glpPQSTV*
Sugars	glycerol 3-phosphate transport	*upgaACE*
Sugars	multiple sugar transporter	*malK, mltK*
Sugars	56annitol synthesis	*mtlD*
Sugars	trehalose synthesis	*otsAB* *trePYZ*
Sugars	proline synthesis	*proAB*
Ectoine	ectoine transport	*DctP* *ehuABCD*
Ectoine	ectoine synthesis	*ectABCD*

Methylamine	choline/carnitine/betaine/glycine transporter	*TCBCT* *betLPTS* *opuABC*
Methylamine	glycine-betaine/proline transport	*proVWX*
Methylamine	glutamine transport	*glnHPQ*
Methylamine	glycine-betaine synthesis	*betAB*

Table 2.3. Osmoprotection genes and categorization for **Figure 2.9**.

Chapter 2 References

1. Anovitz, L. M. & Cole, D. R. Characterization and analysis of porosity and pore structures. *Rev. Mineral. Geochem.* **80**, 61–164 (2015).
2. Mouser, P. J., Borton, M., Darrah, T. H., Hartsock, A. & Wrighton, K. C. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol. Ecol.* **92**, fiw166 (2016).
3. Davies, R. J., Mathias, S. A., Moss, J., Hustoft, S. & Newport, L. Hydraulic fractures: How far can they go? *Mar. Pet. Geol.* **37**, 1–6 (2012).
4. Liang, R. *et al.* Metabolic Capability of a Predominant Halanaerobium sp. In Hydraulically Fractured Gas Wells and Its Implication in Pipeline Corrosion. *Front. Microbiol.* **7**, (2016).
5. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).
6. Struchtemeyer, C. G., Davis, J. P. & Elshahed, M. S. Influence of the drilling mud formulation process on the bacterial communities in thermogenic natural gas wells of the Barnett Shale. *Appl. Environ. Microbiol.* **77**, 4744–4753 (2011).
7. Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K. & Mouser, P. J. Temporal Changes in Microbial Ecology and Geochemistry in Produced Water from Hydraulically Fractured Marcellus Shale Gas Wells. *Environ. Sci. Technol.* **48**, 6508–6517 (2014).
8. Engelder, T. Capillary tension and imbibition sequester frack fluid in Marcellus gas shale. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3625 (2012).
9. Electricity explained: Electricity in the United States. *US Department of Energy, Washington, DC* https://www.eia.gov/energyexplained/electricity/electricity-in-the-us.php (2020).
10. J Johnson, R. J., Folwell, B. D., Wirekoh, A., Frenzel, M. & Skovhus, T. L. Reservoir Souring – Latest developments for application and mitigation. *J. Biotechnol.* **256**, 57–67 (2017).
11. Cliffe, L. *et al.* Identification of Persistent Sulfidogenic Bacteria in Shale Gas Produced Waters. *Front. Microbiol.* **11**, 1–13 (2020).

12. Gaspar, J. *et al.* Microbial Dynamics and Control in Shale Gas Production. *Environ. Sci. Technol. Lett.* **1**, 465–473 (2014).
13. Booker, A. E. *et al.* Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* **2**, e00257-17 (2017).
14. Akob, D. M., Cozzarelli, I. M., Dunlap, D. S., Rowan, E. L. & Lorah, M. M. Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Appl. Geochem.* **60**, 116–125 (2015).
15. Mohan, A. M., Bibby, K. J., Lipus, D., Hammack, R. W. & Gregory, K. B. The functional potential of microbial communities in hydraulic fracturing source water and produced water from natural gas extraction characterized by metagenomic sequencing. *PloS ONE* **9**, (2014).
16. Murali Mohan, A. *et al.* Microbial community changes in hydraulic fracturing fluids and produced water from shale gas extraction. *Environ. Sci. Technol.* **47**, 13141–13150 (2013).
17. Lipus, D. *et al.* Microbial communities in Bakken region produced water. *FEMS Microbiol. Lett.* **365**, 1–11 (2018).
18. Wang, H., Lu, L., Chen, X., Bian, Y. & Ren, Z. J. Geochemical and microbial characterizations of flowback and produced water in three shale oil and gas plays in the central and western United States. *Water Res.* **164**, (2019).
19. Tinker, K. *et al.* Geochemistry and Microbiology Predict Environmental Niches With Conditions Favoring Potential Microbial Activity in the Bakken Shale. *Front. Microbiol.* **11**, 1–14 (2020).
20. Davis, J. P., Struchtemeyer, C. G. & Elshahed, M. S. Bacterial Communities Associated with Production Facilities of Two Newly Drilled Thermogenic Natural Gas Wells in the Barnett Shale (Texas, USA). *Microb. Ecol.* **64**, 942–954 (2012).
21. Hull, N. M., Rosenblum, J. S., Robertson, C. E., Harris, J. K. & Linden, K. G. Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. *Sci. Total Environ.* **644**, 183–192 (2018).
22. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6585–E6594 (2018).

23. Borton, M. A. *et al.* Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale. *Environ. Microbiol.* **20**, 4596–4611 (2018).
24. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
25. Oren, A. Life at High Salt Concentrations. In *The Prokaryotes* (eds. Rosenberg, E., DeLong, E., Lory, S., Stackebrandt, E. & Thompson, F.) 421–440 (Springer Berlin Heidelberg, 2013).
26. Oren, A. Thermodynamic limits to microbial life at high salt concentrations. *Environ. Microbiol.* **13**, 1908–1923 (2011).
27. Oren, A. Bioenergetic Aspects of Halophilism. *Microbiol. Mol. Biol. Rev.* **63**, 334–348 (1999).
28. Shaffer, D. L. *et al.* Desalination and reuse of high-salinity shale gas produced water: Drivers, technologies, and future directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
29. Gallardo, J. & Blackwell, D. D. Thermal structure of the Anadarko basin. *AAPG Bull.* **83**, 333–361 (1999).
30. Hickey, D. A. & Singer, G. A. Genomic and proteomic adaptations to growth at high temperature. *Genome Biol.* **5**, 117 (2004).
31. Stetter, K. O. Extremophiles and their adaptation to hot environments. *FEBS Lett.* **452**, 22–25 (1999).
32. Abousleiman, Y. N. *et al.* Geomechanics Field and Laboratory Characterization of the Woodford Shale: The Next Gas Play . Preprint at https://doi.org/10.2118/110120-MS (2007).
33. Callahan, B. *et al.* DADA2: High resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 4–5 (2016).
34. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

36. Kang, D. D. *et al.* MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, (2019).
37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
38. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
39. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. Drep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
40. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
41. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* 1–18 (2020) doi:10.1093/nar/gkaa621.
42. Brushnell, B. BBTools software package. Sourceforge.net/projects/bbmap/ (2014).
43. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, 1–20 (2015).
44. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
45. Lefkowitz, E. J. *et al.* Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **46**, D708–D717 (2018).
46. Bolduc, B. *et al.* vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **2017**, 1–26 (2017).
47. Bland, C. *et al.* CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 1–8 (2007).
48. Craciun, S., Marks, J. A. & Balskus, E. P. Characterization of choline trimethylamine-lyase expands the chemistry of glycyl radical enzymes. *ACS Chem. Biol.* **9**, 1408–1413 (2014).
49. Segata, N. *et al.* Metagenomic Biomarker Discovery and Explanation. (2011).

50. Shen, H. & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* **99**, 1015–1034 (2008).
51. Cao, K.-A., Rossouw, D., Robert-Granié, C. & Besse, P. A Sparse PLS for Variable Selection when Integrating Omics Data. *Stat. Appl. Genet. Mol. Biol.* **7**, 1544–6115 (2008).
52. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
53. Takahata, Y., Nishijima, M., Hoaki, T. & Maruyama, T. Thermotoga petrophila sp. Nov. and Thermotoga naphthophila sp. Nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan. *Int. J. Syst. Evol. Microbiol.* **51**, 1901–1909 (2001).
54. Chhabra, S. R., Shockley, K. R., Ward, D. E. & Kelly, R. M. Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium Thermotoga maritima grown on glucan- and mannan-based polysaccharides. *Appl. Environ. Microbiol.* **68**, 545–554 (2002).
55. Balk, M., Weijma, J. & Stams, A. J. M. Thermotoga lettingae sp. Nov., a novel thermophilic, methanol-degrading bacterium isolated from a thermophilic anaerobic reactor. *Int. J. Syst. Evol. Microbiol.* **52**, 1361–1368 (2002).
56. Huber, R. *et al.* Thermotoga maritima sp. Nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Arch. Microbiol.* **144**, 324–333 (1986).
57. Stuckman, M. Y., Lopano, C. L., Berry, S. M. & Hakala, J. A. Geochemical solid characterization of drill cuttings, core and drilling mud from Marcellus Shale Energy development. *J. Nat. Gas Sci. Eng.* **68**, 102922 (2019).
58. Evans, M. V. *et al.* In situ transformation of ethoxylate and glycol surfactants by shale-colonizing microorganisms during hydraulic fracturing. *ISME J.* **13**, 2690–2700 (2019).
59. Barati, R. & Liang, J. T. A review of fracturing fluid systems used for hydraulic fracturing of oil and gas wells. *J. Appl. Polym. Sci.* **131**, 1–11 (2014).
60. Reynolds, M. A. A technical playbook for chemicals and additives used in the hydraulic fracturing of shales. *Energy Fuels* (2020) doi:10.1021/acs.energyfuels.0c02527.
61. Welch, S. A. *et al.* Comparative geochemistry of flowback chemistry from the Utica/Point Pleasant and Marcellus formations. *Chem. Geol.* **564**, 120041 (2021).

62. Koskella, B. & Meaden, S. Understanding bacteriophage specificity in natural microbial communities. *Viruses* **5**, 806–823 (2013).
63. Díaz-Muñoz, S. L. & Koskella, B. *Bacteria-Phage interactions in natural environments*. *Advances in Applied Microbiology* vol. 89 (Elsevier Inc., 2014).
64. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**, 2–9 (2012).
65. Holmfeldt, K. *et al.* The Fennoscandian Shield deep terrestrial virosphere suggests slow motion 'boom and burst' cycles. *Commun. Biol.* Doi:10.1038/s42003-021-01810-1.
66. V engosh, A. *et al.* The Geochemistry of Hydraulic Fracturing Fluids. *Procedia Earth Planet. Sci.* **17**, 21–24 (2017).
67. Shaffer, D. L. *et al.* Desalination and reuse of high-salinity shale gas produced water: Drivers, technologies, and future directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
68. Booker, A. E. *et al.* Deepsubsurface pressure stimulates metabolic plasticity in shale-colonizing Halanaerobium spp. *Appl. Environ. Microbiol.* **85**, 1–16 (2019).
69. Nixon, S. L. *et al.* Genome-Resolved Metagenomics Extends the Environmental Distribution of the Verrucomicrobia Phylum to the Deep Terrestrial Subsurface . *mSphere* **4**, 1–18 (2019).
70. Frock, A. D., Gray, S. R. & Kelly, R. M. Hyperthermophilic Thermotoga species differ with respect to specific carbohydrate transporters and glycoside hydrolases. *Appl. Environ. Microbiol.* **78**, 1978–1986 (2012).
71. Wang, Z. *et al.* The Temperature Dependent Proteomic Analysis of Thermotoga maritima. *PloS ONE* **7**, (2012).
72. Cray, J. A. *et al.* The biology of habitat dominance; can microbes behave as weeds? *Microb. Biotechnol.* **6**, 453–492 (2013).
73. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
74. Singer, E. *et al.* Genomic Potential of Marinobacter aquaeolei, a Biogeochemical "Opportunitroph". *Appl. Environ. Microbiol.* **77**, 2763–2771 (2011).

Chapter 3: Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales

3.1 Summary

Viruses are the most ubiquitous biological entities on earth. Even so, elucidating the impact of viruses on microbial communities and associated ecosystem processes often requires identification of unambiguous host-virus linkages – an undeniable challenge in many ecosystems. Subsurface fractured shales present a unique opportunity to first make these strong linkages via spacers in CRISPR-Cas arrays and subsequently reveal complex long-term host-virus dynamics. Here, we sampled two replicated sets of fractured shale wells for nearly 800 days, resulting in 78 metagenomes from temporal sampling of six wells in the Denver-Julesburg Basin (Colorado, USA). At the community level, there was strong evidence for CRISPR-Cas defense systems being used through time and likely in response to viral interactions. Within our host genomes, represented by 202 unique MAGs, we also saw CRISPR-Cas systems were widely encoded. Together, spacers from host CRISPR loci facilitated 2,110 CRISPR-based viral linkages across 90 host MAGs spanning 25 phyla. We observed less redundancy in host-viral linkages and fewer spacers associated with hosts from the older, more established wells, possibly reflecting enrichment of more beneficial spacers through time. Leveraging temporal patterns of host-virus linkages across differing well ages, we report how host-virus co-existence dynamics develop and converge through time, possibly reflecting selection for viruses that can evade host CRISPR-Cas systems. Together, our findings shed light on the complexities of host-virus interactions as well as long term dynamics of CRISPR-Cas defense amongst diverse microbial populations.

---

[1] This chapter was reproduced verbatim from "Amundson, et al. Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales. Current Biology (2023)". The text benefitted from writing and editing contributions from contributing authors and reviewers selected by the publisher. The ordering of the materials in this dissertation are consistent with the content available online but have been renumbered to reflect incorporation into this dissertation.

3.2 Introduction

Viruses are abundant and important constituents of microbial communities in nearly all ecosystems. Consequently, bacteria and archaea, like all living things, are subject to near constant threat of viral predation. In response, many bacteria (~40-60%) and archaea (~90%) deploy CRISPR-Cas viral defense systems[1–4]. CRISPR-Cas works by recording memories of viral interactions via integration of small pieces of viral DNA ('spacers') within the hosts' CRISPR array, that are interspaced with identical repeat sequences and flanked by Cas (CRISPR-associated) genes[5–13]. These saved memories help to protect the host against recurrent invasion by the same viral population by more rapidly identifying and degrading the invading nucleic acids, analogous to antibodies in the human immune system[5–7,10,12,14].

Spacers within CRISPR arrays therefore provide a record of past interactions between a host and viral population, and host-viral linkages can be made by matching the hosts' CRISPR spacers to protospacers in viral genomes[11,15–27]. However, the presence of CRISPR-Cas systems within the microbial community is often a limiting step to making strong host-virus connections; CRISPR-Cas defense is most likely advantageous in ecosystems where host and viral populations repeatedly interact, such as environments dominated by biofilms or those hosting lower microbial and viral diversity[28–30]. Additionally, CRISPR-Cas has been shown to be more widespread in some ecosystems relative to others, such as anoxic environments or those with elevated temperatures[14,28,31–33].

Despite the important role of CRISPR-Cas in viral defense, much remains to be understood about CRISPR-Cas frequency, size, and how presence of these defense systems might influence the temporal dynamics of host and viral populations within diverse microbial

communities. Successful incorporation of a spacer should provide the host future defense upon interaction with the same viral population. However, there are many factors that may influence CRISPR-Cas defense function. For example, CRISPR arrays do not grow exponentially and spacers can indeed be lost,[34–37], and host-viral co-existence despite CRISPR-Cas defense has been observed[38]. Additionally, spacers nearest to the leading end of the CRISPR arrays are most likely to be effective, as they typically represent more recent viral interactions with less time for mutations to occur within the viral protospacer, although recombination can also influence CRISPR array architecture[8]. Thus, it has been hypothesized and shown in laboratory experiments that select spacers may be more favorably retained if they target evolutionarily conserved portions of the viral genome, providing more effective long-term viral defense[13,39,39]. Although ecosystem resources and genome size are not necessarily limiting factors to array size[40,41], other studies have modeled the optimum CRISPR cassette size based on other factors, such as viral diversity and tradeoffs between Cas machinery and array size[42–44]. As a result, it has been suggested that maintaining smaller arrays, on the order of a few dozen to a hundred spacers, may be the optimal size for CRISPR arrays that provide broad protection against a range of viruses but do not overwhelm CRISPR-Cas machinery[42,44,45]. However, many of these insights are derived from modeling or laboratory experiments and there remains a need to understand patterns of CRISPR-Cas defense in environmental systems with diverse microbial communities.

To address this knowledge gap, we used a temporally-resolved dataset from six fractured shale wells to interrogate host-virus dynamics and CRISPR-Cas loci in a subsurface ecosystem. Subsurface fractured shales, which are relatively closed ecosystems with limited immigration, elevated temperatures, lower microbial diversity and likely dominated by biofilms, present an opportunity to address these questions through strong CRISPR-based host-viral linkages[21,23,46–49].

We hypothesized and found that CRISPR-Cas viral defense systems were widely encoded across hosts within shale microbial communities. Building on this, we applied multiple bioinformatic approaches to identify CRISPR spacers in both recovered host genomes and metagenomes and made strong host-virus linkages for many of the recovered host genomes. This approach also facilitated temporal investigations into host utilization of CRISPR-Cas at the community-level and host-population levels to better understand CRISPR-Cas defense in this natural ecosystem. Finally, we leveraged over two thousand viral linkages to investigate host-viral dynamics and saw evidence for increase host-virus co-existence through time. To our knowledge, our study represents one of the most extensive analyses of long-term, host-viral temporal dynamics with CRISPR-based linkages in an environmental system to date.

3.3 Results & Discussion

3.3.1 Fractured shale ecosystems provide a unique opportunity to investigate virus-host temporal dynamics.

We sampled fluids from two sets of hydraulically fractured oil & gas wells in the Denver-Julesburg (DJ) Basin for nearly 800 days (Colorado, USA). The two sets of wells were defined by their age relative to the initial fracturing process: the 'established' wells operated for nearly three years prior to the initiation of our sampling campaign (DJB-1, DJB-2, DJB-3), while we began sampling the 'new' wells shortly after they had been hydraulically fractured (DJB-4, DJB-5, DJB-6) (**Figure 3.1**). All three wells within each group were located on the same frack pad and subject to the same drilling and hydraulic fracturing process, resulting in three replicate wells for each group.

From a total of 78 metagenomes across all six wells (**Figure 3.1 & Appendix C**), we recovered 202 unique metagenome assembled genomes (MAGs) representing 29 phyla and 2,176 unique viral MAGs (vMAGs) >10kb from the subsurface communities. The microorganisms that persist in this ecosystem likely originate from water, sand, and chemical inputs used during the hydraulic fracturing process. Many of the dominant and persisting MAGs – encompassing bacterial taxa affiliated with *Clostridia*, *Thermotogae*, *Fusobacteriia*, and *Synergistia* and archaeal taxa affiliated with *Methanosarcinia*, *Methanomicrobia*, and *Thermococci* (**Figure 3.2**) – have been reported in other engineered subsurface environments[21,50–54], and their relative abundances in this system reflected patterns observed in complementary 16S rRNA gene analyses (**Figure 3.2).** Although vMAGs were recovered from metagenomes and not viromes, only a small portion of viruses were predicted to be temperate by presence of integrase genes (*n=192*) or HMM searches of domains associated with temperate viruses (*n=293*)[55].

Microbial communities, however, are not static through time. Taxa unable to tolerate high temperatures and elevated salinity are likely outcompeted, while biofilms and spatially distinct niches likely emerge and expand in this closed ecosystem[56,57]. Thus, we expect that microbial communities within the established wells are more spatially heterogeneous, while microbial communities in the new wells are initially well mixed and more spatially homogenous[49,58]. In agreement with these assumptions, we observed higher host (bacterial and archaeal) and viral alpha diversity in the established wells relative to the new wells (*Wilcox, p=5.563e-06*) (Figure S2). Alpha diversity also generally increased through time in all wells, likely reflecting the development of niches fostering more diverse taxa (**Figure 3.3**). This trend contrasts findings from previous fractured shale studies that reported a rapid decrease in microbial diversity[46,47]. More broadly, host alpha diversity in the DJ Basin was also higher than many other fractured

shale ecosystems studied to date[50,52,59,60] and similar to those reported previously for produced fluids from the DJ Basin[51]. Notably, microbial communities from the new wells became more similar to those in the established wells over time, likely reflecting the maturation of the well ecosystem (**Figure 3.3**). Together, these results illustrate the temporal juxtaposition of the two sets of wells, and the connectedness of host-viral dynamics in increasingly diverse microbial communities within a closed, subsurface ecosystem.

3.3.2 Evidence for active viral predation in deep subsurface shales microbial communities

Although community composition did vary with time, host and viral community dynamics generally mirrored one another (**Figure 3.4**). We quantified these temporal changes in community structure using Bray-Curtis dissimilarity values (**Figure 3.4**). In this analysis, higher dissimilarity values indicate greater change in community composition relative to the previous timepoint. Shifts in host and viral communities were strongly and positively correlated, often mirroring one another in their dynamics (**Figure 3.4**) – a trend which was also reflected in host and viral alpha diversity (*Spearman Rho: old=0.71, 0.96, 0.076 new=0.6, 0.66, 0.91*). Viruses depend on their hosts for replication, yet host populations are often impacted as a direct result of this proliferation. Thus, the strong relationship observed here suggests that host and viral communities were continually changing, and viral predation was likely occurring. Interestingly, we did not observe trends toward community stability in either grouping of wells, which would be indicated by consistently decreasing Bray-Curtis dissimilarity values. Finally, there was a stronger relationship between host and viral communities in the established wells compared to the new wells, potentially reflecting the temporal loss of viral populations that lack hosts and subsequent enrichment of interdependent host and viral populations.

Only a small portion of recovered viruses encoded genes indicative of a temperate lifestyle (**Appendix C**), suggesting that these strong correlations are not driven by integrated prophage. A temperate lifestyle may not be necessary for a virus to survive in this ecosystem, as has been reported for other more extreme environments, due to availability of growth substrates (i.e., organic carbon) from additives used in the fracturing process and reduced environmental stress on microbiomes due to the lower salinity of the DJ Basin (avg. 47ms/cm). Additionally, only a very small proportion of vMAGs that encoded an integrase gene (<5 in each well) closely matched their hosts' coverage, indicating that temperate viruses are unlikely to be solely responsible for the trends observed. Instead, lytic viruses are likely recovered during the filtration step in sample processing as produced fluids are often viscous, containing small particulates that clog the filter pores and elevated levels of ferric iron which can bind viral capsids[23,61]. Thus, the strong relationships observed here are likely due to dynamic virus host interactions and driven by both lytic and lysogenic infections.

3.3.3 Community-level responses to viral interactions recorded by CRISPR-Cas arrays

Viral diversity has been shown to impact the success of, and selection for, CRISPR-Cas systems, as the 'memory' recording of a viral interaction as an integrated spacer is more effective in ecosystems where repeated interactions between host and viral populations occur[29]. The closed nature of these subsurface ecosystems should promote such repeated interactions, and thus we sought to identify evidence of hosts using CRISPR-Cas defense at the community level.

First, repeats and spacers from CRISPR-Cas arrays (which tend to break during metagenomic assembly) were identified in all samples using CRASS[62]. Overall, we recovered a total of 918,724 spacers from all 78 metagenomes. All six wells had a significant positive

relationship between the number of spacers recovered and time, especially in the new wells where the total number of spacers rapidly increased through the first one hundred days (**Figure 3.5**). Although the established wells initially contained greater total number of spacers compared to the new wells, later timepoints in the new wells began to approach established wells' totals, highlighting the speed at which spacers may be incorporated by microorganisms.

Finally, linking this temporal relationship with observations that the viral community is continually changing, we observed that the number of spacers also had a significant and strong positive relationship with the total number of unique viral populations (vMAGs) across all samples (**Figure 3.5**). Together, these results demonstrate the presence of widely encoded CRISPR-Cas defense systems in the microbial community and suggest that microorganisms using CRISPR-Cas are likely responding to ongoing viral interactions and integrating matching spacers into CRISPR arrays through time.

3.3.4 CRISPR-Cas & other viral defense systems within host genomes

Moving beyond community-level analyses, we sought to link individual CRISPR-Cas viral defense systems to representative host MAGs. Our host MAGs, by nature, are composite genomes that likely represent a host population, as opposed to a single host cell. Linking spacers assembled from metagenome reads to CRISPR-Cas loci associated with a given MAG thus provides an overview of the total complement of spacers encoded by members of this population, and a window into the population-level diversity and dynamics at these CRISPR-Cas loci. In total, 123 of our 202 MAGs (~60%) spanning 25 phyla contained a detectable CRISPR array (**Figure 3.6 & Appendix C**). We identified CRISPR-Cas loci in a higher proportion of MAGs from the established wells (67%) relative to the new wells (54%), highlighting the persistent

widespread use of this defense system in an environment where viral predation is likely recurrent. Type I-B was the most common CRISPR-Cas system type (22%) out of all identified and classified, followed by type III-A (17%) (**Appendix C**). The high proportion of hosts containing a CRISPR-Cas array is not unexpected, as CRISPR-Cas systems are reported to be more widely encoded in closed ecosystems, biofilms, and ecosystems with elevated temperatures – three characteristics of fractured subsurface shales.

We next leveraged CRASS to identify additional spacers associated with our host MAGs[62]. Briefly, repeat sequences identified within host arrays were matched to those identified with CRASS and spacers grouped to the CRASS-identified repeat were then associated with the MAG, representing a host population (**Figure 3.1**). This approach identified thousands of additional spacers from metagenomic reads and associated many of these spacers with host genomes from as many timepoints as possible. Importantly, this facilitated the identification of temporal trends in CRISPR-Cas loci sizes, as insights into host arrays were not limited to a single timepoint where the host MAG was recovered.

For many host populations, the number of CRISPR spacers generally exhibited a strong positive relationship with MAG coverage, a proxy for relative abundance in the community (**Figure 3.7**). This trend was obscured when MAGs were grouped together at higher taxonomic levels (**Figure 3.7**), highlighting the importance of genome-resolved analyses and the need for even more resolved analyses, possibly at the single cell level, to better understand loss and gain of spacers through time in natural communities. In contrast, there was greater variability in the relationship between spacers and time at the genome level (**Figure 3.7**), with few host populations exhibiting consistent increases or decreases in number of spacers over time. The overall increase in number of CRISPR spacers through time observed at the community levels

(**Figure 3.4**) is thus probably not only due to increases in spacer number within individual populations but also to an overall increase in host populations encoding active CRISPR-Cas systems.

CRISPR-Cas is one of many viral defense systems, many of which have been only recently described[63]. Not all MAGs encoded a detectable CRISPR array, and thus we hypothesized that other viral defense systems are likely deployed by hosts. We found that 87% of all MAGs contained another viral defense system, with no significant difference in the proportion of MAGs between the new and established wells (**Appendix C**). Most of the MAGs that contained a CRISPR array also tended to encode another defense system in both the new and established wells (81% and 78% of MAGs, respectively). However, there was a greater diversity of different viral defense systems encoded in the established wells compared to the new wells, with 41 and 34 different systems detected across MAGs from the established and new wells, respectively[63]. In all wells the most common viral defense system was a restriction modification system, which works more promiscuously than CRISPR-Cas to degrade nucleic acids[64], while known abortive systems that induce cell death (i.e., Abi2, AbiEii) were a small proportion of the overall number of systems in both established (6%) and new (4%) wells[65,66]. Together, this provides further evidence that there is a benefit to the host for encoding viral defense mechanism(s) and offers insights into the types of defense systems that may be paired with CRISPR-Cas in an environmental system.

3.3.5 Fewer CRISPR spacers associated with hosts in the established wells may reflect selection towards more effective CRISPR-Cas arrays

With continued interactions between hosts and viruses, we hypothesized that MAGs (representing host populations) in the established wells would generally be associated with more CRISPR spacers reflecting these events. However, loss of spacers through time as well as a theoretical optimum for CRISPR array size[34–36,42,44] have both been previously reported, which could lead to populations from established wells encoding fewer spacers. Here, despite higher host and viral diversity in the established wells, we observed on average a greater number of spacers per host population in the new wells (avg. 288±410), compared to the established wells (avg. 180±306) (**Figure 3.8**). The high number of spacers associated with hosts may be due to MAGs representing host populations rather than a single host cell, and likely mirrors trends within the subsurface host populations and reflects differences in terms of population diversity amongst the group of wells for CRISPR loci.

A greater number of viral linkages were also made per representative host MAG from the new wells relative to the established wells, averaging 22 and 7 unique viral linkages, respectively (**Figure 3.8**). We hypothesize that fewer linkages per host in the established wells may be driven by interactions with fewer different viruses, potentially due to more heterogeneous and confined spatial structures where host and viral populations interact. Although MAGs from the established wells encoded fewer spacers and linked to less viruses, we observed less redundancy within those viral linkages. For all linkages made (i.e., every host linked to any virus) for MAGs from the established wells, an average of 71% of those linkages were to unique viruses (**Figure 3.9**). In the new wells we observed greater redundancy in linkages to the same virus, as only 39% of all linkages were to unique viruses. Therefore, while MAGs from the established wells contained

fewer spacers and linked to less viruses, they could be matched to a proportionally greater number of different viruses, suggesting that the retained spacers may help to protect the host against a wider suite of viruses with less redundancy.

One way host populations with less spacers may still provide efficient defense against viral predation is if the retained spacers target regions of the viral genome with fewer mutations. We used single nucleotide polymorphism (SNP) frequency as a proxy for sequence variation in viral genes to investigate possible spacer effectiveness. Overall, very few spacers persisted in both sets of wells, highlighting the continual fluctuations in host and viral communities and the loss and gain of spacers within hosts populations. More spacers from the established wells persisted (4.5%) for at least half the sampling timepoints compared to the new wells (1.3%). Further, we observed that spacers recovered at only one timepoint generally matched viral protospacers with the highest SNP frequency (**Figure 3.10**). Additionally, the percentage of targeted viral genes with zero SNPs increased with spacer persistence in the new wells, but not the established wells (**Figure 3.10**). That is, spacers that were present across the most timepoints tended to target viral genes that had less sequence variation within the community. These trends may be associated with the increased selection and retention of spacers that may confer viral resistance for longer periods of time.

3.3.6 Temporal increase in patterns of host-virus co-existence

Spacers within CRISPR-Cas arrays can be uniquely leveraged to make strong inferences about host-virus dynamics, as spacers from the host array often identically match the viral protospacer 'target'[15]. Leveraging additional spacers identified via CRASS, we were able to identify 2,110 viral linkages across 90 MAGs representing 25 different phyla (**Figure 3.6**).

Indeed, matching all spacers associated with a host MAG to our vMAG database yielded at least one viral linkage for a majority of MAGs encoding a CRISPR-Cas array. We observed ≥1 viral linkage for 68% of MAGs with CRISPR-Cas arrays in the established wells (48 of 70 MAGs) and 79% of MAGs with CRISPR-Cas arrays in the new wells (42 of 53 MAGs) (**Figure 3.6, Appendix C**). There was no significant relationship between MAG coverage and the number of viral linkages in the established wells, and a weak positive relationship in the new wells (Spearman's rho *R=0.39, p=1.4e-07*), suggesting that there is not a sustained relationship between the hosts' overall success and the number of different interacting viruses.

Even amongst widespread CRISPR-Cas defense and the presence of matching (linking) spacers, many linked viruses persisted. Therefore, we next evaluated how often viruses can persist and interact with the host population, despite theoretical CRISPR-Cas defense. We leveraged our 2,110 host-virus linkages to quantify differences in host-virus co-occurrence patterns in both sets of wells and studied how these dynamics may develop through time. We quantified occurrences of three scenarios for every host-virus linkage (1) when only the host was present, (2) when both the virus and host were present, and (3) when only the virus was present (Figure 6). Scenarios where both host and virus were absent were excluded. In this analysis, 'absence' of host or virus is likely not complete absence of the virus in the ecosystem, but rather indicates that their true abundance was very low and there was insignificant evidence for their presence.

We observed differing patterns of host-virus dynamics between the new and established wells, potentially reflecting the establishment of microbial communities through time in this closed ecosystem. Notably, we tracked a decreasing trend in occurrences of only the host present in the new wells, approaching values observed in the established wells. Concurrently, we

observed a slight increasing trend in host-virus co-existence in both sets of wells. This provides evidence that CRISPR-Cas may be most effective when microbial communities are first introduced into the newly formed ecosystem. CRISPR-Cas may then become less effective through time as the selection of viruses able to evade host defenses results in greater frequency of host-virus co-existence (**Figure 3.11**).

Anti-CRISPR (*acr*) genes are one mechanism employed by viruses to persist despite host defense, as they can suppress CRISPR-Cas systems. Putative anti-CRISPR genes were identified in 16 different vMAGs that were persistent in the established wells. Although this is a small proportion of total vMAGs, *acr* genes are poorly characterized and infrequently observed in natural ecosystems, and thus likely under-detected in this dataset. Together, these findings shed light on the complexities of host-virus dynamics temporally and how subsurface closed ecosystems may develop towards an equilibrium of host-virus co-existence, as opposed to dominance by host or viral populations[67] or 'red queen' dynamics of constant evolution and population turnover[68,69].

3.4 Conclusions

Here we leveraged time-resolved samples from six hydraulically fractured shale wells to establish CRISPR-based host-viral linkages and study long-term host-virus co-existence and CRISPR-Cas dynamics in a natural, closed ecosystem. Timeseries data (>800 days) from all six wells allowed us to recover CRISPR spacers from metagenomes and MAGs, which facilitated community-level and host-population level analyses of CRISPR-Cas defense through time. At the community level, we observed evidence that viral predation is active through time and that hosts are likely incorporating new spacers into their arrays in response to viral interactions. Next,

at the genome-level, we observed that CRISPR-Cas viral defense systems were widely encoded across a majority of MAGs. In total, we identified CRISPR arrays in ~60% of MAGs across 25 of 29 different phyla representing host populations from the deep shale microbial communities. We observed that host populations (represented by MAGs) in the established wells were associated with fewer spacers, and that there was less redundancy in viral linkages, potentially reflecting selection for retention of more effective spacers through time in a closed ecosystem.

Leveraging CRISPR spacers to link viruses to hosts, we next identified potential viruses for a majority of hosts containing a CRISPR-Cas array, with over two thousand total linkages identified across 90 different host MAGs. The proliferation of microenvironments (e.g., biofilms) over time in these subsurface ecosystems may constrain the number of interactions between diverse host and viral populations, resulting in fewer linkages in the established wells. Alternatively, such patterns may be attributed to lack of viral recovery due to successful host defense. Finally, given the prevalence of CRISPR-Cas systems and the important role this defense might have in host-viral co-existence, we used host-viral linkages to interrogate host-virus temporal dynamics. We found that co-existence of host and viral populations generally increased through time, potentially due to the selection for viruses able to evade host defenses, specifically CRISPR-Cas defense, in this closed ecosystem. Together, this study offers new insights into the long-term dynamics between host and viral populations and CRISPR-based host-viral linkages within a subsurface ecosystem.

3.5 Materials & Methods

3.5.1 Experimental Model and Subject Details

Produced fluid samples were collected from six hydraulically fractured wells from the Niobrara formation, within the Denver-Julesburg (DJ) Basin, in eastern Colorado between October 2018 and October 2020 (*n=78*) (**Figure 3.1**). The Niobrara shale formation consists of three benches that are located approximately 1890-1950 meters deep in the subsurface with a downhole temperature measuring approximately 112°C (recorded while drilling). The six wells within this formation were sampled are split equally into two groups defined by their age when sample collection began: the three 'established' wells (*n=33*) had been producing for approximately 1000 days prior to sample collection (DJB-1, DJB-2, and DJB-3), while the three 'new' wells (*n=45*) were sampled from day ~30 in production (DJB-4, DJB-5, and DJB-6). A small number of early produced fluid samples (those beginning with 'JMDJ#', <60 days) were collected directly from well heads and filtered through a 0.22µm pore size polyethersulfone membrane Sterivex filter due to field sampling constraints (MilliporeSigma) with a minimum of 500mL of fluid filtered. Most produced fluids (those beginning with 'DJKA') were collected directly from separator tanks into 1L Nalgene bottles with no head space and stored at 4°C until processing, which occurred within 24 hours from when the sample was collected. To the degree possible, samples were collected from separator tanks shortly after the last contents had been released to the central processing facility. 500-800mL of fluid was filtered through a 0.2µm PES membrane Nalgene vacuum filtration unit (Thermo Scientific). Filters were removed from the units and stored at -20°C until DNA extraction. Therefore, MAGs were recovered from produced fluids collected from the separator tank or well head for each well, though for brevity we refer to

these MAGs as simply recovered from the well. Conductivity was measured on raw, unfiltered fluids at room temperature using a Myron L 6PIIFCE meter.

3.5.2 DNA extraction and metagenomic sequencing

Total nucleic acids were extracted from half of each sample's 0.2μm filter using DNAeasy PowerSoil Kit (Qiagen). Extraction blanks were run with each round of DNA extractions and all returned no detectable nucleic acids using the maximum amount of blank sample (20μL) via the Qubit dsDNA High Sensitivity assay kit (ThermoFisher Scientific). For all 78 samples, genomic DNA was prepared for metagenomic sequencing at the Genomics and Microarray Core at the University of Colorado, Denver's Genomics Shared Resource. Samples were prepared using the Illumina Nextera XT Library System according to manufacturer's instructions for 2x150bp libraries and were sequenced using the Illumina NovaSeq platform and paired-end reads were collected.

3.5.3 16S rRNA gene sequencing and analysis

Nucleic acids for all samples were also sent to Argonne National Laboratory for 16S rRNA gene sequencing (**Appendix C**). Sequencing was performed with the Illumina MiSeq platform, using the Earth Microbiome Project primer set for amplification of the 251bp hyper-variable V4 region. 16S rRNA gene sequences were obtained via Argonne's standard procedure, with the exception of performing 30 PCR amplification cycles. Paired-end reads were processed with QIIME2 (v 2021.2) EMP protocol, by first demultiplexing via exact-matching of barcodes, trimmed to 250bp and denoised with DADA2[70]. Representative sequences were taxonomically classified with SILVA (release 138). 16S community composition results are shown in **Figure**

3.2. All 16S rRNA gene sequencing reads were submitted to NCBI under BioProject PRJNA308326 and individual BioSample accession numbers are listed in **Appendix C**.

3.5.4 Metagenomic assembly, binning, and viral recovery

For bacterial, archaeal, and viral recovery, total sequenced DNA from each sample was first trimmed from 5' to 3' ends with Sickle (https://github.com/najoshi/sickle) and individually assembled using IDBA-UD with default parameters[71]. Assembly information for each sample is provided in **Appendix C**. Only scaffolds ≥5kb from metagenomic assemblies were used for binning bacterial and archaeal genomes with MetaBAT2 (v2.12.1) to recover metagenome assembled genomes (MAGs)[72]. CheckM (v.1.1.2) lineage workflow ('lineage_wf') followed by the 'qa' command was used to assess completion and contamination for each metagenomic bin[73], and medium (>50% completion, <10% contamination) and high (>90% completion, <5% contamination) quality bins were recovered from all samples from all six wells following the standard metrics for MAGs proposed by Bowers et al.[74]. The two sets of unique MAGs (from the new and established wells) were individually determined by dRep v2.2.3 using default parameters[75]. MAGs were dereplicated based on their well groupings so that representative host populations were most reflective of true host populations in the subsurface communities, and to identify host repeats and associate spacers from CRASS. We anticipate differences such as the age differences (including possible differences in additive used), as well as physical separation of the well groupings from one another could impact host genomic content, specifically repeats in CRISPR arrays, and thus we created a MAG database unique to each grouping of wells. We refer to the final set of 202 MAGs as the 'host' community (Figure S1). All MAGs were taxonomically classified using GTDB-Tk v2.2.0[76]. Metagenomic assemblies and MAGs were

annotated via DRAM (v1.2.4) using default parameters[77]. Additional details about MAGs can be found in **Appendix C**. Metagenomic reads and MAGs were submitted to NCBI BioProject PRJNA308326 and individual accession numbers can be found in **Appendix C**.

Viral MAGs (vMAGs) were also identified in metagenomic assemblies from scaffolds ≥10kb in length using VirSorter2 (v2.2.2)[78] and following the "Viral sequence identification SOP with VirSorter2" developed by the Sullivan Lab[79]. Following this protocol, quality of vMAGs were assessed using checkV (v0.8.1) and annotated using DRAM-v (v1.2.4)[77,80]. Low confidence vMAGs were removed following the manual curation steps in the SOP. Viral genomic contigs (≥10kb) were clustered into viral populations (genus level) using the 'ClusterGenomes' (v 1.1.3) app in CyVerse using the parameters 95% average nucleotide identity and 90% alignment fraction of the smallest contig (https://github.com/simroux/ClusterGenomes). The resulting database of 2,176 vMAGs are considered our viral database. Viral taxonomy was determined by clustering vMAGs with viruses belonging to the viral reference taxonomy databases in NCBI Bacterial and Archaeal Viral RefSeq v211, and viruses from the International Committee on Taxonomy of Viruses (ICTV) via vConTACT2 v0.11.3 with default settings[81]. Anti-CRISPR (arc) genes were identified in vMAGs using ArcFinder (using both homology-based and guilt by association based approaches) with default parameters[82]. Probable viral lifestyle (either lytic or lysogenic) was inferred via one of two methods: (1) presence of integrase genes via KEGG annotation and (2) >75% confidence of a temperate lifestyle assigned from BACPHLIP (HMM searching for temperate domains)[55]. All vMAGs have been deposited under NCBI BioProject PRJNA308326 and additional details about vMAGs can be found in **Appendix C**.

3.5.5 Calculating MAG and vMAG coverage and relative abundance

To calculate coverage and relative abundance of MAGs and vMAGs, all 78 pairs of trimmed metagenomics reads were rarified to the lowest metagenome sequencing depth of 9Gbp using the 'reformat' guide within bbmap[83]. Coverage for MAGs was calculated by competitively mapping rarified metagenomic reads to MAGs using bbmap (v38.89) with minid=90. Resulting sam files were converted to sorted bam files using samtools (v1.9)[84]. Coverage for each MAG was calculated using coverM (v0.6.0) (https://github.com/wwood/CoverM) using two commands. First, coverM was run using –min-covered-fraction=90 to determine MAGs read recruitment to at least 90% of the genome. Second, coverage values were calculated using the -m reads_per_base command, which represents reads mapped/genome length, and thus multiplied this by read length (151bp) in order to calculate MAG coverage (simply, coverage = reads_per_base * 151 bp). Only MAGs with >1x coverage and with reads mapped to >90% of the genome were considered present in a sample. Relative abundance was thus calculated as the proportion of a given MAG's coverage out of the sum of all present MAGs' coverage, per sample.

Metagenomic reads were also mapped to vMAGs to determine coverage using bbmap with minid=95 (v38.89)[83] and sam files converted to bam files using samtools (v1.9)[84]. Given vMAGs are viral contigs, coverM (https://github.com/wwood/CoverM) contig mode was applied with two commands. First, --min-covered-fraction 75 and next followed by -m reads_per_base to calculate coverage. Similar to requirements set for MAGs, here vMAGs must have a minimum covered fraction >75% to be considered present. Coverage values were calculated from the reads per base output*151 bp. Number of viruses present in a metagenome were determined by presence of vMAGs given this recruitment of metagenomic reads.

3.5.6 Detection of viral defense systems and recovery of spacers

CRISPR arrays in MAGs were identified using the Geneious (v.2020.0.5) plugin CRISPR Recognition Tool (CRT)[85] v.1.2 using the 'Find CRISPR loci' annotation tool with the following parameters: min number of repeats a CRISPR must contain: 4, minimum length of a CRISPR's repeated region: 19, maximum length of a CRISPR's repeated region: 55, minimum length of a CRISPR's non-repeated region (or spacer region): 19, maximum length of a CRISPR's non-repeated region (or spacer region), length of a search window used to discover CRISPR's: 8. CRISPR arrays were then classified into types/subtypes using CRISPRCasTyper (v.1.8.0)[86] via matching repeat sequences. Spacers were also detected in non-rarified and rarified trimmed metagenomics reads using CRisprASSembler: CRASS (v1.0.1)[62]. Briefly, CRASS reassembles CRISPR-Cas arrays of repeats and spacers that tend to break during assemblies and groups spacers by the repeat sequences in CRISPR arrays. Only spacers recovered from rarified metagenomics reads were used to represent the 'total number of spacers in a metagenome' for all community-level correlations to not introduce bias from varying read depth. All recovered host genomes, regardless of detection of a CRISPR array, were also queried for 60 other known anti-phage systems using DefenseFinder (v.1.0)[63].

3.5.7 Making CRISPR-based host-virus linkages

Linkages between MAGs and vMAGs (hosts and viruses) were made exclusively via CRISPR spacers using two approaches (**Figure 3.1**). As a result of this, linkages could only be made with MAGs that had a detectable CRISPR array. First, CRISPR arrays were identified in MAGs using Geneious, and spacers and repeats were extracted from the CRISPR arrays. We

then leveraged CRASS to make as many linkages as possible and evaluate the number of spacers associated with a MAG through time. Repeat sequences from MAGs were identically matched to direct repeat sequences from CRASS (same length, no mismatches). Spacers that were associated with a direct repeat sequence from CRASS were thus grouped with the MAG of the same repeat sequence. To make as many host-viral linkages as possible, spacers were extracted from CRASS applied to non-rarified reads. Next, spacers from all MAGs (linked via Geneious and CRASS) were queried against all vMAGs using BLASTn with the parameters to optimize short sequences BLAST: -dust no and -word_size 7. Finally, only identical or nearly identical (0 or 1 mismatch across spacer length) were used to match spacers to vMAGs and make host-viral linkages. Number of linkages per MAG is shown in **Figure 3.8**.

3.5.8 Host-viral co-occurrence patterns

All MAGs with at least one viral linkage were included in analyzing host-viral co-occurrence patterns. For each MAG and individual linked virus at every timepoint, all possibilities were evaluated for being one of three interaction types: (1) only host present but virus absent (below detection), (2) both host and virus are present and (3) when only the virus was present, but their linked host was absent (below detection). Instances where both host and virus were not present were excluded from any calculations and not counted in the total number of interaction occurrences, which was used to normalize occurrences. Thus, the percent of any interaction was calculated as the proportion of all interactions previously stated.

3.5.9 Analysis of single nucleotide polymorphisms in vMAGs

We combined SNP values for viral genes with the persistence of spacers that link host and virus to interrogate any possible relationship between gene variation and spacer retention for all MAGs with linkages (**Figure 3.10**). We utilized MetaPop[87] with default parameters to calculate the number of SNPs within all viral genes identified in our vMAGs. For genes that met MetaPop's default parameters, SNP frequency was calculated relative to the gene length. Genes containing linked protospacers that did not meet MetaPop's default parameters were not included in this analysis. Finally, SNP frequency for the gene containing the protospacer was combined with the persistence of the spacer (i.e., number of samples the spacer was recovered).

3.5.10 Quantification and Statistical analysis

Alpha diversity (Shannon's index) and beta diversity (Bray-Curtis) values were calculated using vegan v2.6-2 in R. Alpha diversity was calculated using 16S rRNA amplicon data, while beta diversity and Bray-Curtis dissimilarity values were calculated based on the host and viral communities recovered via metagenomic sequencing and rarified reads. Metagenomics was used here since the viral community was recovered using metagenomics and thus the paired host communities were assessed similarly via MAGs (recovered from metagenomes). Bray-Curtis dissimilarity values were calculated as the difference in beta diversity from the previous timepoint. Spearman correlations and p-values were calculated using ggpubr to determine the strength and directionality of relationships between variables such as number of spacers, MAG/vMAG coverage, time, etc. Specifically, correlations between number of spacers and host coverage were only calculated for MAGs that were both present in at least 3 timepoints and also had spacer recovery from at least three timepoints.

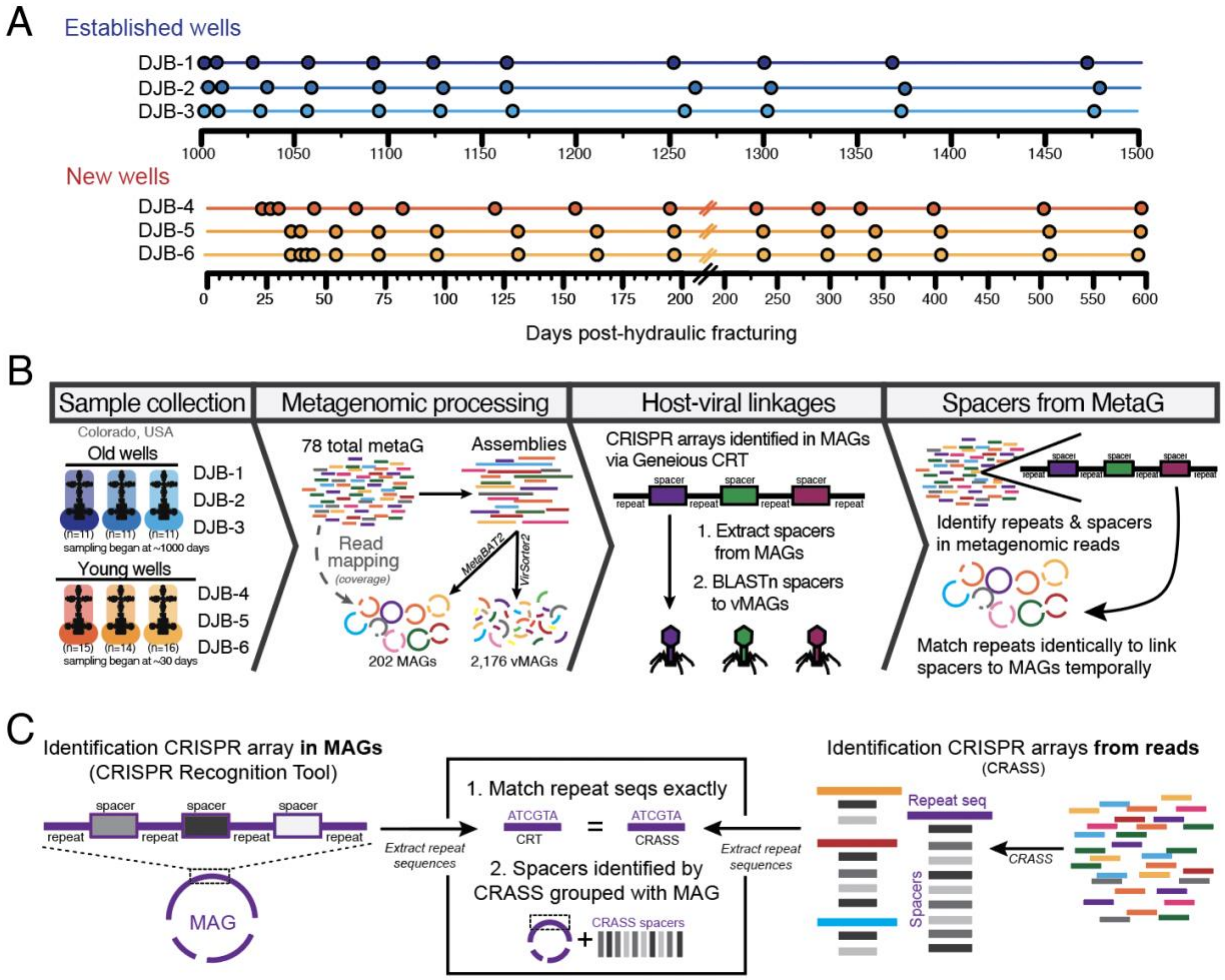

Figure 3.1. Sampling scheme and methods. **(A)** Sampling timepoints for all six wells split by their age group. Dashes at 200 days on the axis of the new wells sampling scheme represents a change in scale. **(B)** Overview of methods used to recover representative host & viral genomes, make linkages between MAGs & vMAGs, and identifying spacers from metagenomic reads for a community level insight. **(C)** Overview of methods used to link additional spacers, identified with CRASS, to representative host genomes.

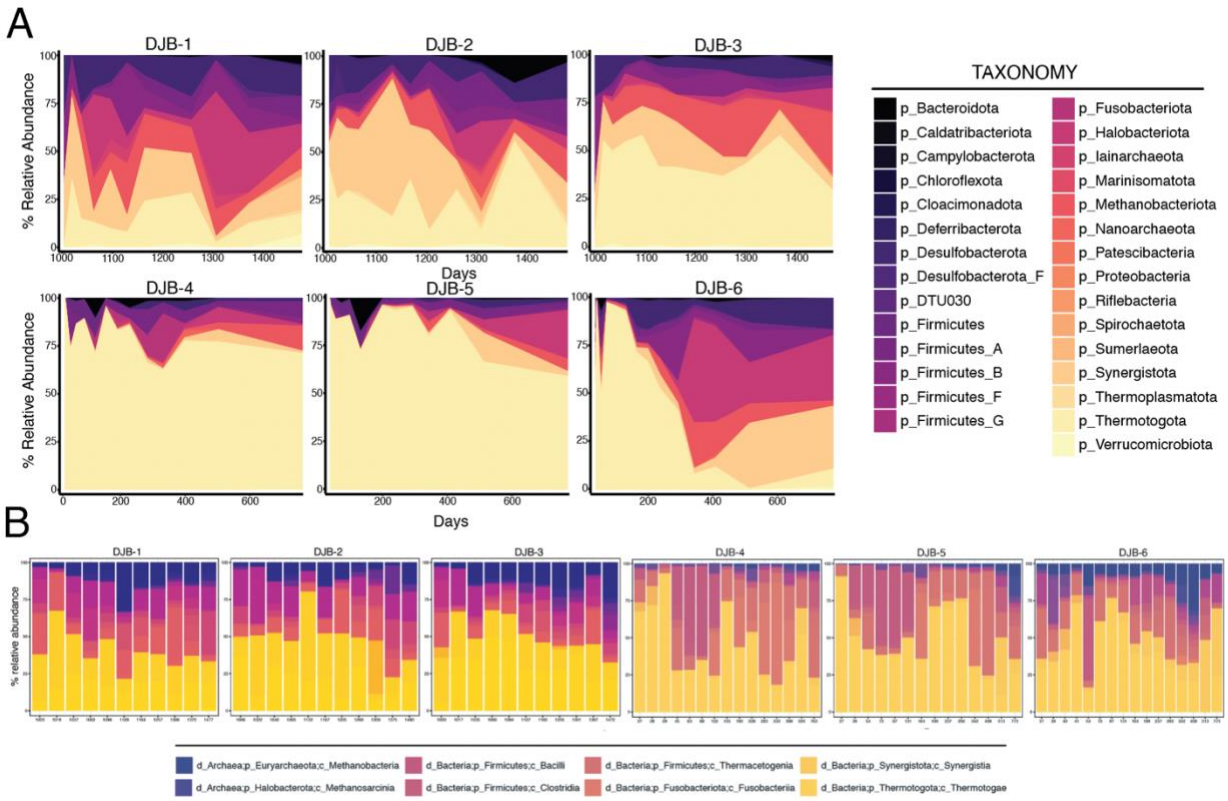

Figure 3.2. Community composition of all six wells. **(A)** Temporal dynamics of the microbial community as shown by MAGs relative abundance in all six wells, summed and colored at the phyla level. **(B)** Barcharts illustrating the temporal dynamics of DJ Basin microbial communities based on16S rRNA gene amplicon data. Select dominant taxa that were observed in 3+ wells are highlighted below bar charts.

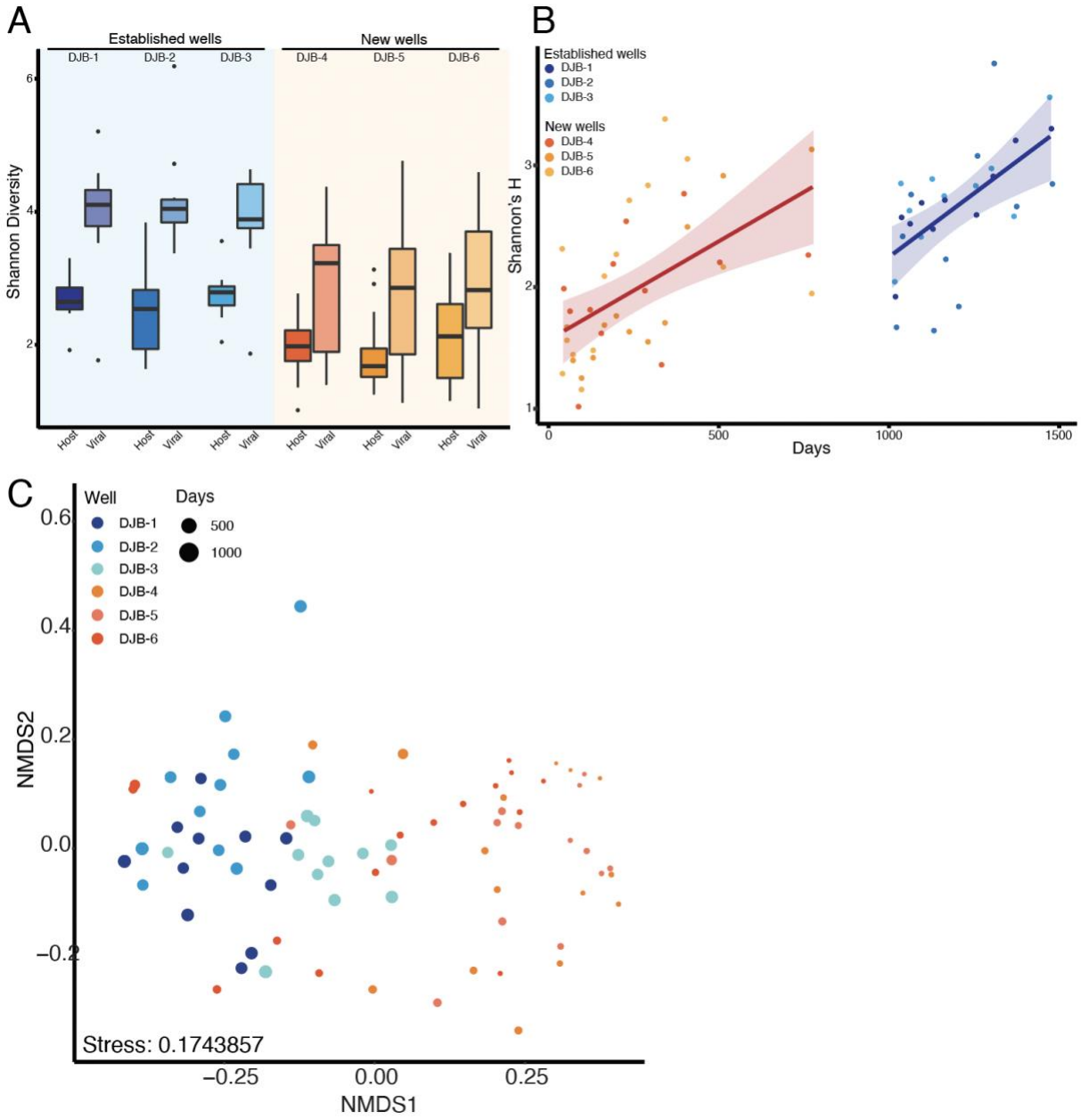

Figure 3.3. Alpha and beta community diversity metrics. **(A)** Boxplots of alpha diversity (Shannon's) for host and viral diversity for all six wells. **(B)** Temporal alpha diversity of 16S rRNA gene amplicon data (Shannons') through time reveals a temporal increase in bacterial & archaeal alpha diversity across all six wells. **(C)** NMDS ordination of complimenting 16S rRNA gene amplicon data using Bray-Curtis distance.

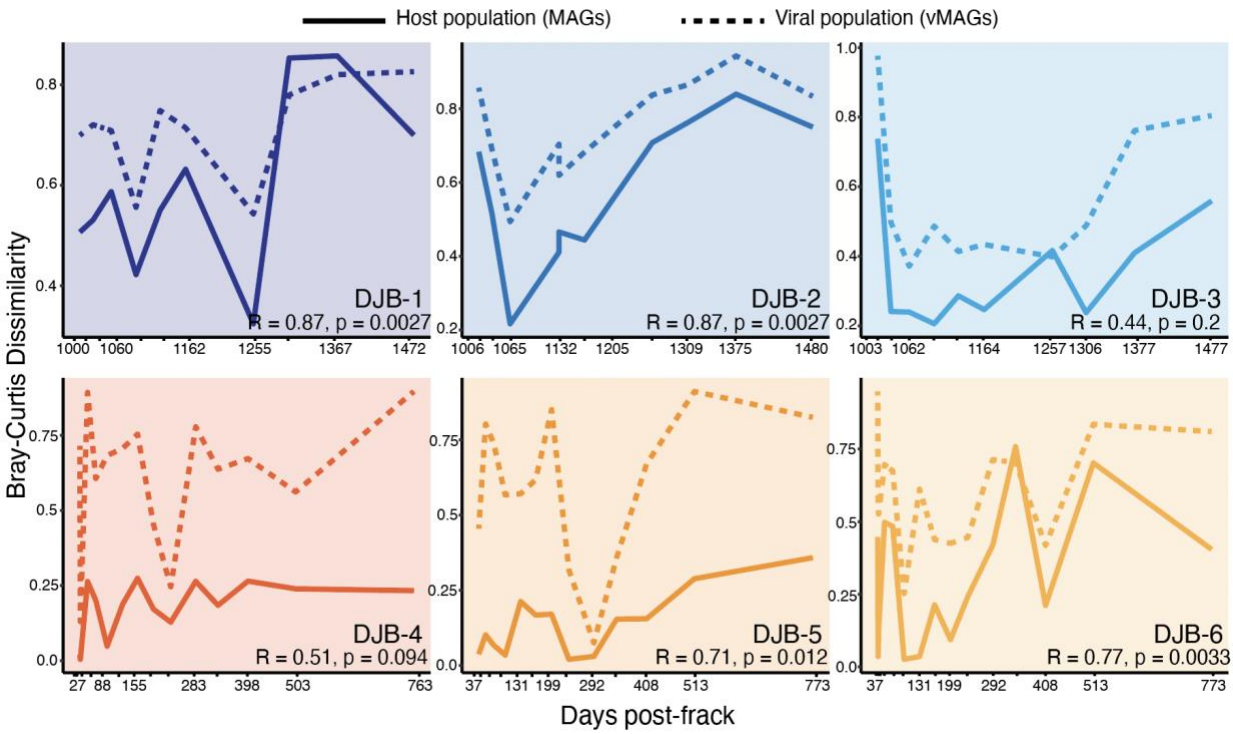

Figure 3.4. Temporal dynamics of host (bacterial and archaeal MAGs) and viral (vMAGs) communities. Bray-Curtis community dissimilarity through time for both host (solid line) and viral (dashed line) communities illustrating the change in each community composition from the previous timepoint, with larger dissimilarity values indicating greater change in community structure. Spearman's rho and p values highlight the strong positive relationship between the temporal changes in host and viral populations in all wells.

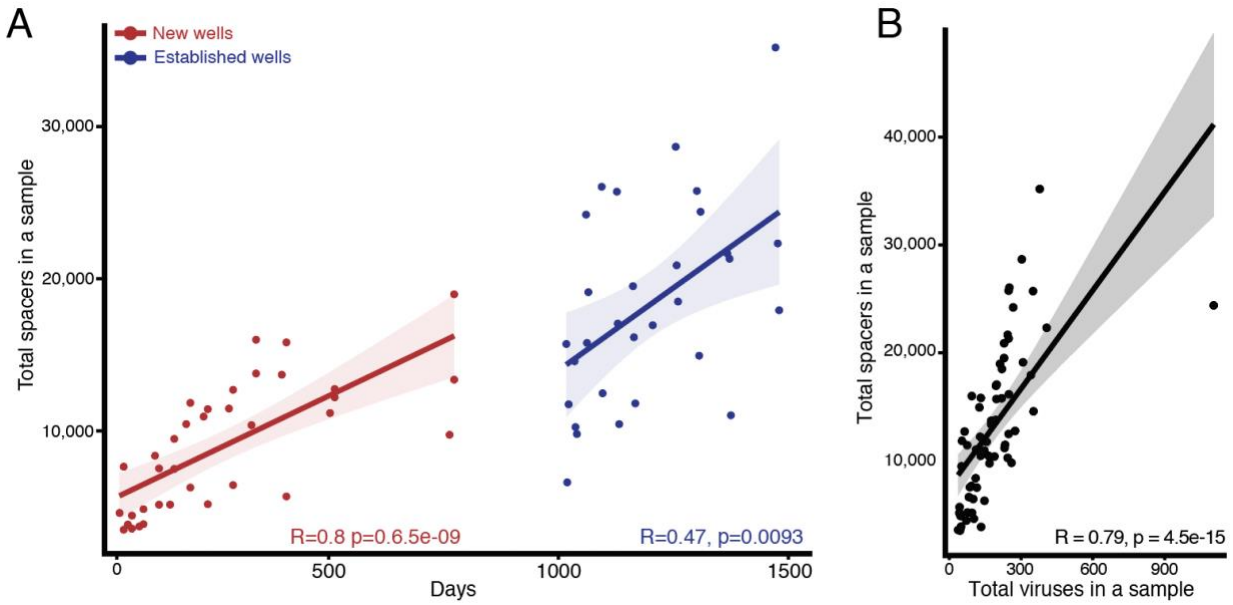

Figure 3.5. Community-level responses to viral interactions through time as recorded by CRISPR arrays within the microbial community. **(A)** Spearman correlation between the number of spacers recovered in a sample and the number of viruses in a sample. **(B)** Spearman correlations between the number of spacers in a sample and days for the new and established wells.

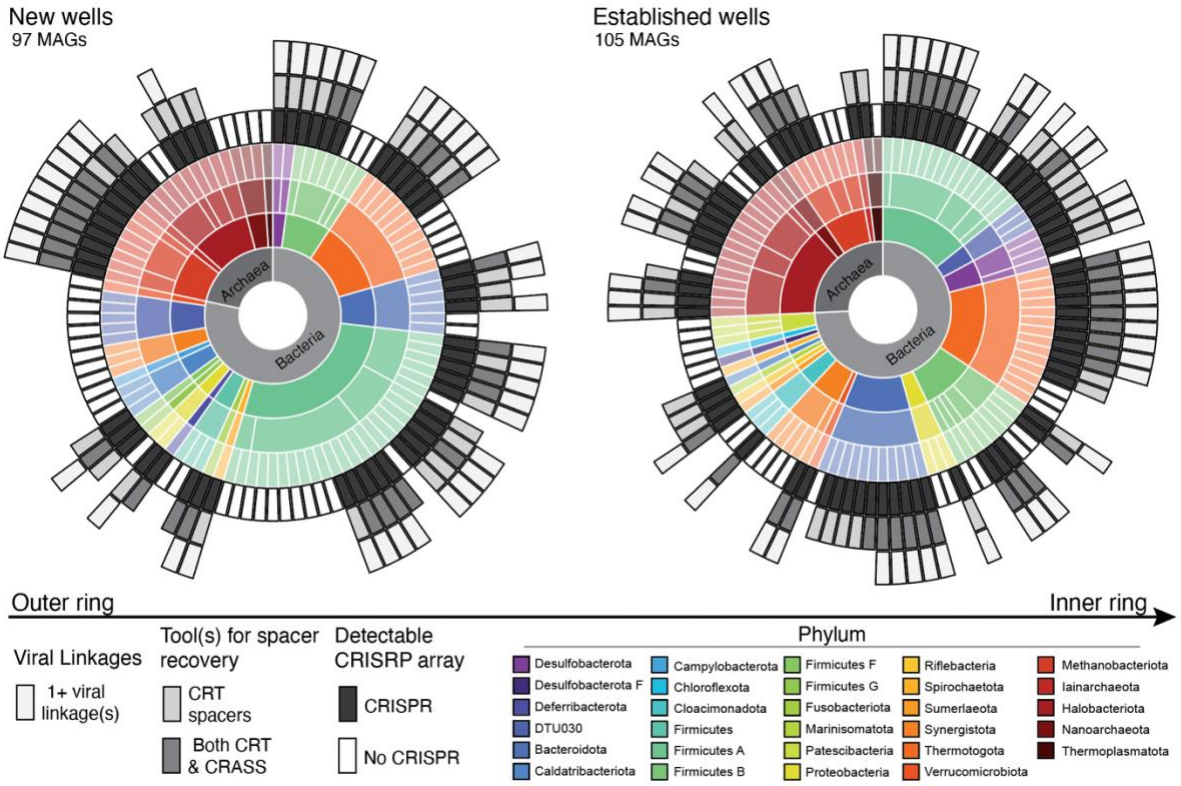

Figure 3.6. CRISPR arrays and viral linkages in representative host genomes (MAGs) recovered in the new and established wells, organized by phylum. The first four rings (from inside out) provide information on the taxonomy of host MAGs and how many MAGs were recovered for a given phyla. Inner most ring splits MAGs by bacteria or archaea. The second ring illustrates different phyla represented. The third ring illustrates how many different taxonomic classes are represented by MAGs from a given phyla. And the fourth ring shows how many individual host MAGs were recovered for a given phyla & class. The last three rings present information on CRISPR arrays (presence/absence), how spacers were recovered (just CRISPR Recognition Tool or CRT and CRASS), and if at least one viral linkage was made for the MAG.

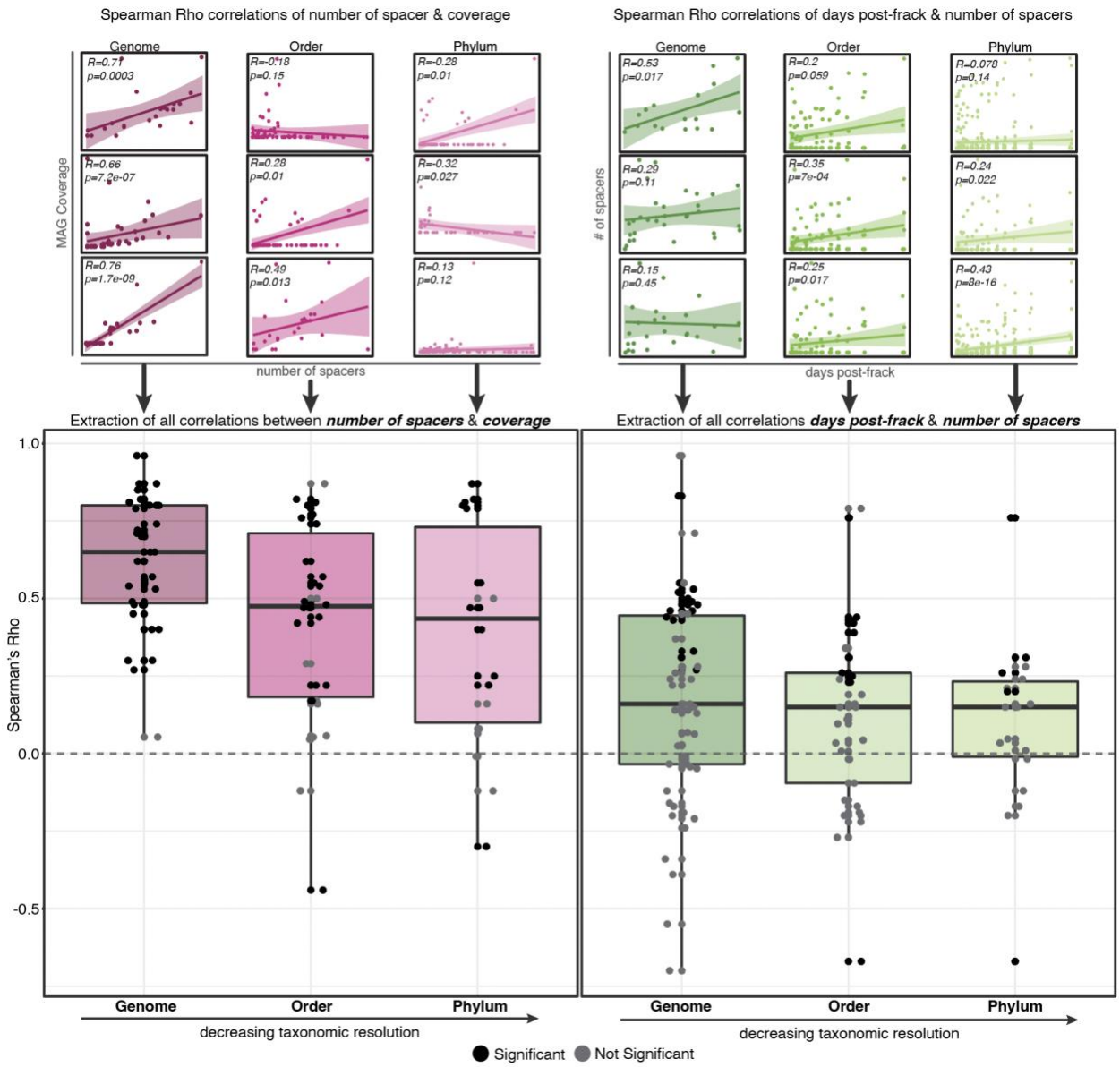

Figure 3.7. Correlations between number of spacers associated with a MAG and MAG coverage or days post frack at three decreasing levels of taxonomic resolution. Top panels illustrate the generation of data (Spearman's rho) that is represented in the boxplots below, with three examples provided per taxonomic level. All rho values and p values for each individual genome, order, or phyla were extracted and plotted as an individual data in the boxplots below for both variables: coverage (pink) or days post-frack (green). Non-significant (>0.05) correlation values (grey) are shown to highlight the lack of a significant positive or negative relationship in many cases, especially at higher taxonomic levels, which is also highlighted in examples above.

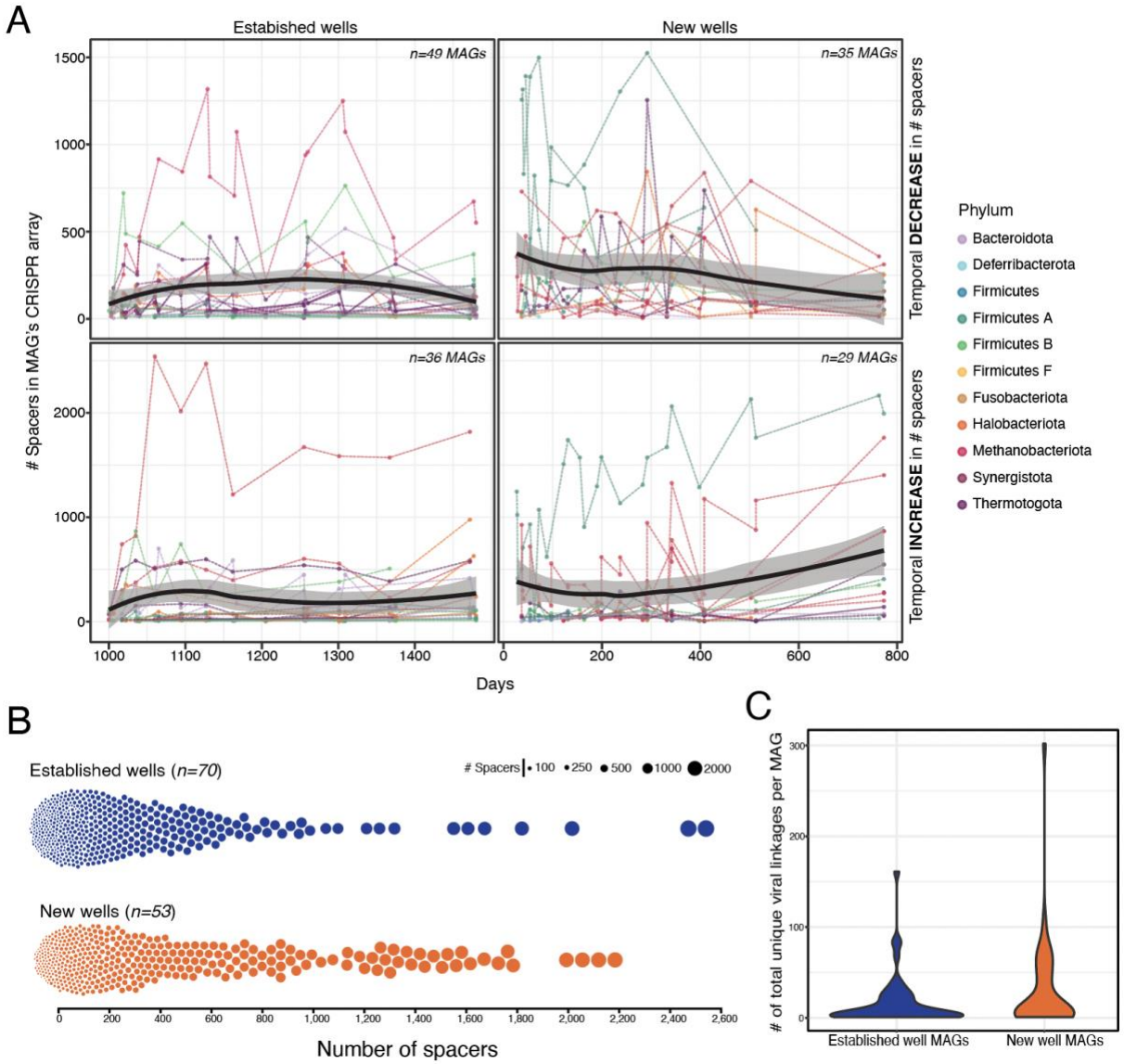

Figure 3.8. (A) Temporal changes in the number of spacers in a MAG CRISPR array. Each dotted line represents a single MAG with general trends highlighted as solid black lines. Top panels show MAGs with arrays that generally decreased (final timepoint < average array size) for the established (left) and new (right) wells, while bottom panels show MAGs with arrays that generally increased (final timepoint > average array size). In general, MAG arrays were smaller and fluctuated less in the established wells than in the new wells. **(B)** Range in the number of spacers in MAG CRISPR arrays. Each dot is both sized and arranged by the size of the CRISPR arrays, colored by new or established wells. Additionally, each dot represents one MAG at one timepoint, and therefore may be multiple points for MAGs that had spacers recovered at alternate timepoints via CRASS. **(C)** Distribution of total number of linkages, per MAG, to unique vMAGs. Linkages were made using paired CRASS recovered spacers as well as Geneious identified spacers and therefore viral linkages may have been made via spacers recovered at many different timepoints, but distribution shows the total per MAG.

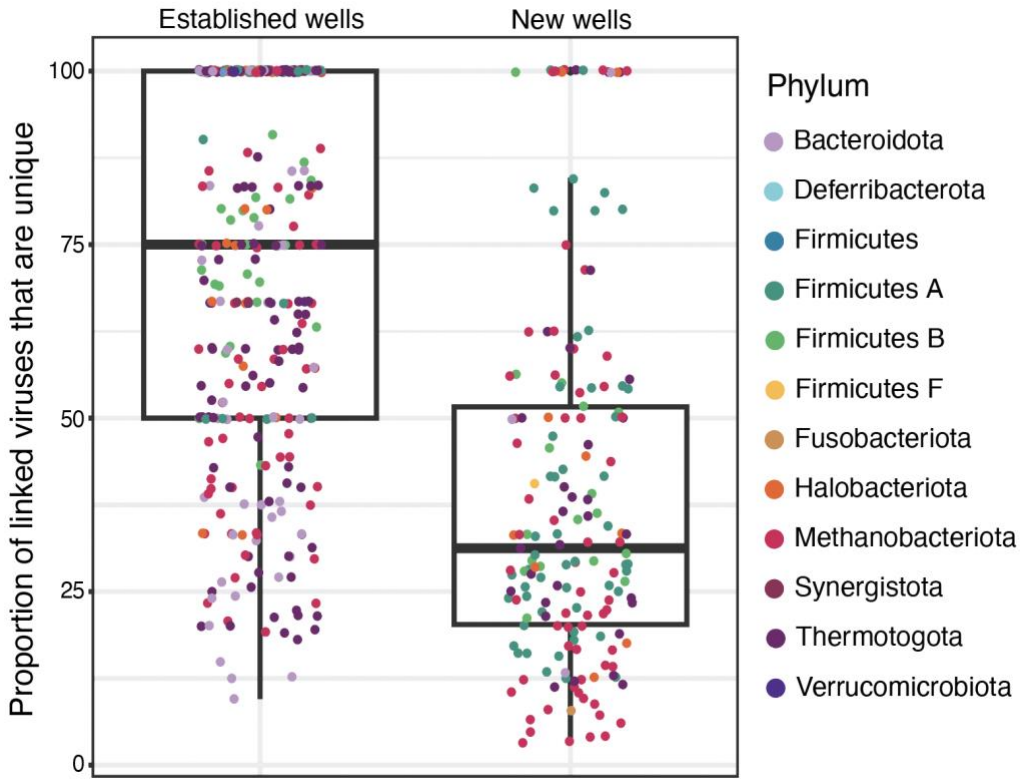

Figure 3.9. Boxplots illustrating the proportion of linked viruses, per MAG, that are unique. Each point represents a single host population at a single timepoint where spacers were recovered and linkages were made, colored by phylum. The proportion of unique viruses linked was calculated as the number of different viruses linked to out of the total number of linkages at a given timepoint.

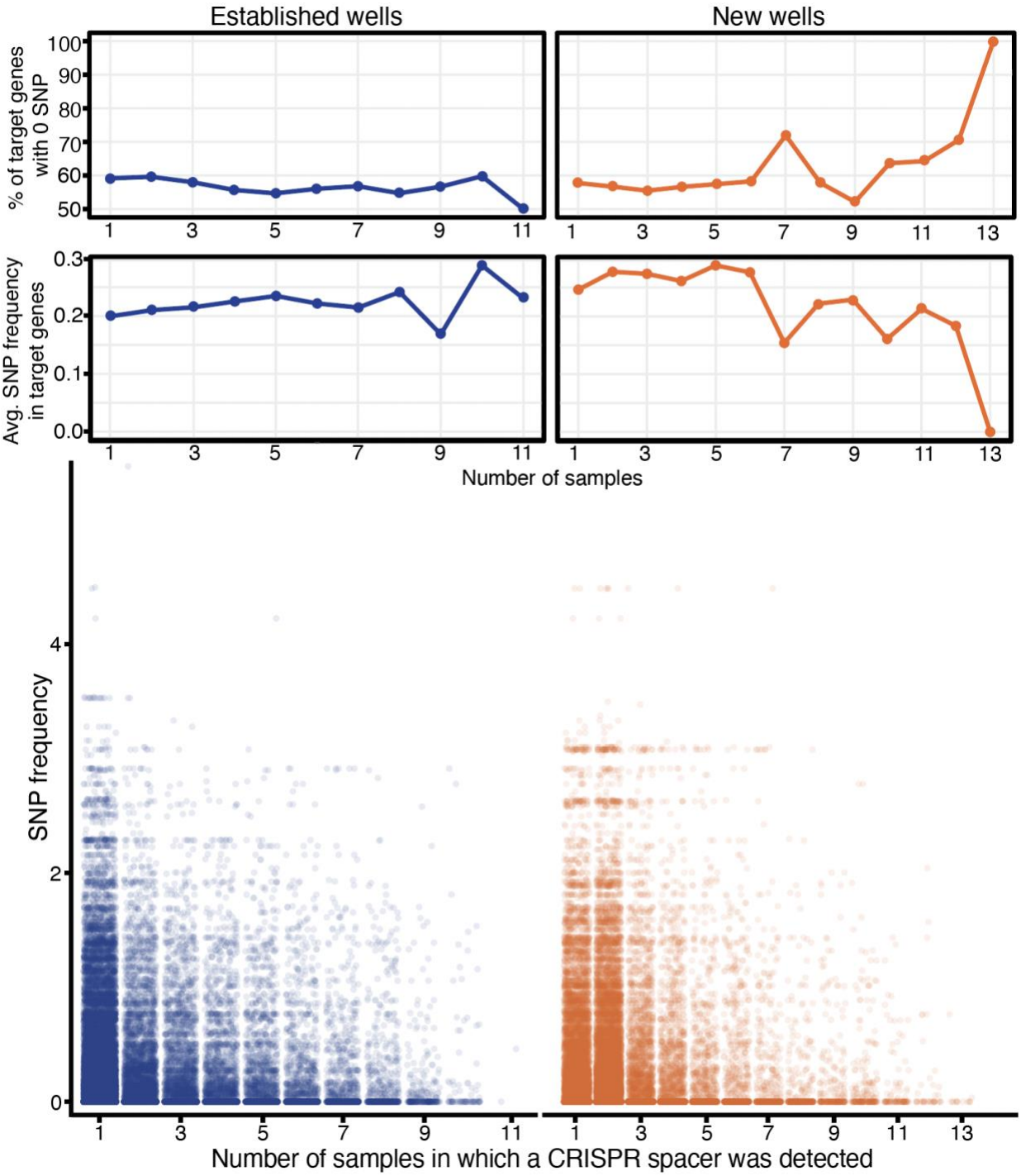

Figure 3.10. SNP frequency (number of SNPs/gene length) for protospacer genes with corresponding persistence of the host spacer. Spacer persistence is quantified as the number of samples that the spacer was recovered. Average SNP frequency as well as percent of viral genes with zero SNPs is shown as line graphs above raw data plots.

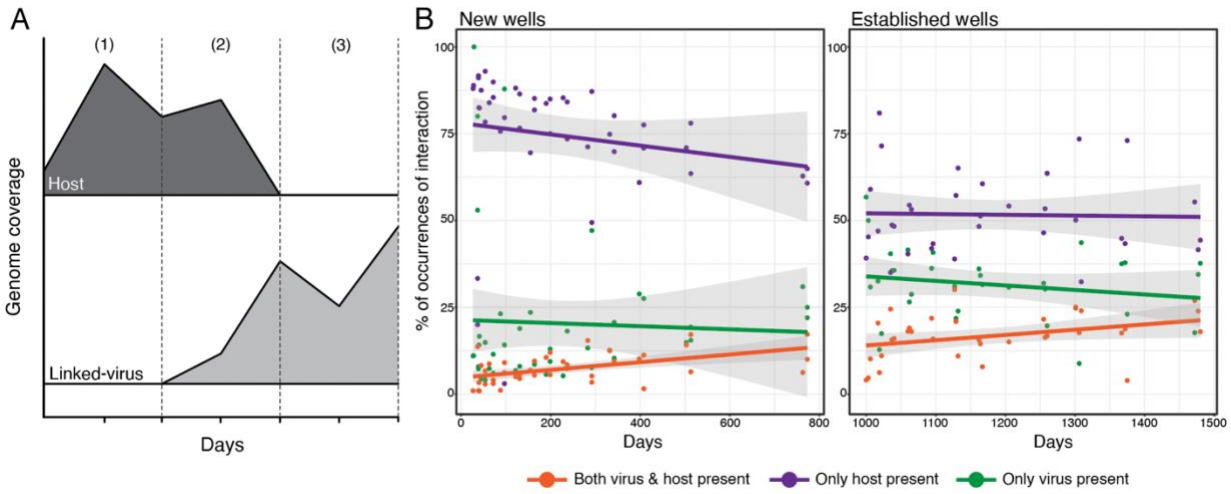

Figure 3.11. Patterns of host-virus co-existence. **(A)** Conceptual diagram of different 'interaction' types: (1) where only host is present, but the virus was not present (below detection) (2), where host & virus are both present and (3) where only the virus is present. Axes are purposefully left blank given for this conceptual illustration. **(B)** Temporal trends in percent of each interaction type in the new and established wells. Lines represent linear trends while shaded grey areas indicate the 95% confidence interval.

Chapter 3 References

1. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
2. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).
3. Staals, R. H. J. & Brouns, S. J. J. Distribution and Mechanism of the Type I CRISPR-Cas Systems. in *CRISPR-Cas Systems: RNA-mediated Adaptive Immunity in Bacteria and Archaea* (eds. Barrangou, R. & van der Oost, J.) 145–169 (Springer, 2013). doi:10.1007/978-3-642-34657-6_6.
4. Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).
5. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
6. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).

7.Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, eaal5056 (2017).

8. Hille, F. *et al.* The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**, 1239–1259 (2018).
9. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20180087 (2019).
10. Barrangou, R. & Marraffini, L. A. CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Mol. Cell* **54**, 234–244 (2014).
11. Andersson, A. F. & Banfield, J. F. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science* **320**, 1047–1050 (2008).
12. Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* **327**, 167–170 (2010).
13. Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1401–1412 (2008).
14. Watson, B. N. J., Steens, J. A., Staals, R. H. J., Westra, E. R. & Houte, S. van. Coevolution between bacterial CRISPR-Cas systems and their bacteriophages. *Cell Host Microbe* **29**, 715–725 (2021).

15. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
16. Anderson, R. E., Brazelton, W. J. & Baross, J. A. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol. Ecol.* **77**, 120–133 (2011).
17. Sanguino, L., Franqueville, L., Vogel, T. M. & Larose, C. Linking environmental prokaryotic viruses and their host through CRISPRs. *FEMS Microbiol. Ecol.* **91**, fiv046 (2015).
18. McKay, L. J. *et al.* Sulfur cycling and host-virus interactions in Aquificales-dominated biofilms from Yellowstone's hottest ecosystems. *ISME J.* **16**, 842–855 (2022).
19. Emerson, J. B. *et al.* Virus-Host and CRISPR Dynamics in Archaea-Dominated Hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* **2013**, e370871 (2013).
20. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
21. Amundson, K. K. *et al.* Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. *Microbiome* **10**, 5 (2022).
22. Berg, M. *et al.* Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. *ISME J.* **15**, 1569–1584 (2021).
23. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
24. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).
25. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
26. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci.* **110**, 12450–12455 (2013).
27. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).

28. Weinberger, A. D., Wolf, Y. I., Lobkovsky, A. E., Gilmore, M. S. & Koonin, E. V. Viral Diversity Threshold for Adaptive Immunity in Prokaryotes. *mBio* **3**, e00456-12 (2012).
29. Meaden, S. *et al.* High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. *Curr. Biol.* **32**, 220-227.e5 (2022).
30. Broniewski, J. M., Meaden, S., Paterson, S., Buckling, A. & Westra, E. R. The effect of phage genetic diversity on bacterial resistance evolution. *ISME J.* **14**, 828–836 (2020).
31. Bernheim, A. *et al.* Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. *Nat. Commun.* **8**, 2094 (2017).
32. Westra, E. R., Dowling, A. J., Broniewski, J. M. & van Houte, S. Evolution and Ecology of CRISPR. *Annu. Rev. Ecol. Evol. Syst.* **47**, 307–331 (2016).
33. Weissman, J. L., Laljani, R. M. R., Fagan, W. F. & Johnson, P. L. F. Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *ISME J.* **13**, 2589–2602 (2019).
34. Deveau, H. *et al.* Phage Response to CRISPR-Encoded Resistance in Streptococcus thermophilus. *J. Bacteriol.* **190**, 1390–1400 (2008).
35. Bradde, S., Vucelja, M., Teşileanu, T. & Balasubramanian, V. Dynamics of adaptive immunity against phage in bacterial populations. *PLOS Comput. Biol.* **13**, e1005486 (2017).
36. Garrett, S. C. Pruning and Tending Immune Memories: Spacer Dynamics in the CRISPR Array. *Front. Microbiol.* **12**, (2021).
37. Lopez-Sanchez, M.-J. *et al.* The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. *Mol. Microbiol.* **85**, 1057–1071 (2012).
38. Guerrero, L. D. *et al.* Long-run bacteria-phage coexistence dynamics under natural habitat conditions in an environmental biotechnology system. *ISME J.* **15**, 636–648 (2021).
39. Sun, C. L., Thomas, B. C., Barrangou, R. & Banfield, J. F. Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J.* **10**, 858–870 (2016).
40. Levin, B. R., Moineau, S., Bushman, M. & Barrangou, R. The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet.* **9**, e1003312 (2013).

41. Vale, P. F. *et al.* Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. *Proc. R. Soc. B Biol. Sci.* **282**, 20151270 (2015).
42. Martynov, A., Severinov, K. & Ispolatov, I. Optimal number of spacers in CRISPR arrays. *PLOS Comput. Biol.* **13**, e1005891 (2017).
43. McGinn, J. & Marraffini, L. A. CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol. Cell* **64**, 616–623 (2016).
44. Bradde, S., Nourmohammad, A., Goyal, S. & Balasubramanian, V. The size of the immune repertoire of bacteria. *Proc. Natl. Acad. Sci.* **117**, 5144–5151 (2020).
45. Childs, L. M., Held, N. L., Young, M. J., Whitaker, R. J. & Weitz, J. S. Multiscale Model of Crispr-Induced Coevolutionary Dynamics: Diversification at the Interface of Lamarck and Darwin. *Evolution* **66**, 2015–2029 (2012).
46. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).
47. Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K. & Mouser, P. J. Temporal Changes in Microbial Ecology and Geochemistry in Produced Water from Hydraulically Fractured Marcellus Shale Gas Wells. *Environ. Sci. Technol.* **48**, 6508–6517 (2014).
48. Mouser, P. J., Borton, M., Darrah, T. H., Hartsock, A. & Wrighton, K. C. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol. Ecol.* **92**, fiw166 (2016).
49. Booker, A. E. *et al.* Deep-Subsurface Pressure Stimulates Metabolic Plasticity in Shale-Colonizing Halanaerobium spp. *Appl. Environ. Microbiol.* **85**, e00018-19 (2019).
50. Wang, H., Lu, L., Chen, X., Bian, Y. & Ren, Z. J. Geochemical and microbial characterizations of flowback and produced water in three shale oil and gas plays in the central and western United States. *Water Res.* **164**, 114942 (2019).
51. Hull, N. M., Rosenblum, J. S., Robertson, C. E., Harris, J. K. & Linden, K. G. Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. *Sci. Total Environ.* **644**, 183–192 (2018).
52. Murali Mohan, A. *et al.* Microbial Community Changes in Hydraulic Fracturing Fluids and Produced Water from Shale Gas Extraction. *Environ. Sci. Technol.* **47**, 13141–13150 (2013).

53. Murali Mohan, A., Hartsock, A., Hammack, R. W., Vidic, R. D. & Gregory, K. B. Microbial communities in flowback water impoundments from hydraulic fracturing for recovery of shale gas. *FEMS Microbiol. Ecol.* **86**, 567–580 (2013).
54. Struchtemeyer, C. G. & Elshahed, M. S. Bacterial communities associated with hydraulic fracturing fluids in thermogenic natural gas wells in North Central Texas, USA. *FEMS Microbiol. Ecol.* **81**, 13–25 (2012).
55. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
56. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578–6583 (1998).
57. McMahon, S. & Parnell, J. Weighing the deep continental biosphere. *FEMS Microbiol. Ecol.* **87**, 113–120 (2014).
58. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **17**, 247–260 (2019).
59. Tinker, K. *et al.* Geochemistry and Microbiology Predict Environmental Niches With Conditions Favoring Potential Microbial Activity in the Bakken Shale. *Front. Microbiol.* **11**, (2020).
60. Stemple, B. *et al.* Biogeochemistry of the Antrim Shale Natural Gas Reservoir. *ACS Earth Space Chem.* **5**, 1752–1761 (2021).
61. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011).
62. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res.* **41**, e105 (2013).
63. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
64. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
65. Chopin, M.-C., Chopin, A. & Bidnenko, E. Phage abortive infection in lactococci: variations on a theme. *Curr. Opin. Microbiol.* **8**, 473–479 (2005).

66. Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Res.* **42**, 4590–4605 (2014).
67. van Houte, S. *et al.* The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature* **532**, 385–388 (2016).
68. Stern, A. & Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *BioEssays* **33**, 43–51 (2011).
69. Weitz, J. S. *et al.* Phage–bacteria infection networks. *Trends Microbiol.* **21**, 82–91 (2013).
70. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
71. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
72. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
73. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
74. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
75. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
76. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
77. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
78. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).

79. Viral sequence identification SOP with VirSorter2. https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoyqebg4o/v3.
80. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
81. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
82. Yi, H. *et al.* AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. *Nucleic Acids Res.* **48**, W358–W365 (2020).
83. BBMap download | SourceForge.net. https://sourceforge.net/projects/bbmap/.
84. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
86. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J.* **3**, 462–469 (2020).
87. Gregory, A. C. *et al.* MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome* **10**, 49 (2022).

Chapter 4: Leveraging genomic insights for a biogeographical understanding of fractured shale microbiome function

4.1 Summary

Subsurface shales underly much of North America and are economically important for the recovery of oil & natural gas. However, many microorganisms that persist in shale wells contribute to common production challenges such as bioclogging, corrosion, souring, and scaling. Given that shale formation temperature and salinity of produced waters can vary considerably between geographically distinct basins, there is need to understand and predict spatial and temporal patterns of microbial functional potential in these ecosystems. To investigate this, we performed metagenomic sequencing on 209 water samples collected from 36 hydraulically fractured wells across nine shale basins in North America, as well as two international basins (UK & China). Many of these samples were from 19 wells with timeseries data (n=161), and a majority paired with NMR metabolomic data (n=128), which together revealed significant differences in concentrations of detectable metabolites, taxonomic profiles, and the complete absence of a core microbiome across basins. From these samples we built a comprehensive genomic database for fractured shales containing 978 unique medium- and high-quality MAGs and over 7 million unique genes. We leveraged a genome-resolved approach to uncovering the differences in key metabolisms, such as fermentation and sulfate and thiosulfate reduction, across geographically distinct basins. Next, to uncover spatiotemporal patterns in functional potential across 978 unique metagenomic assembled genomes (MAGs) recovered from shale samples, we developed a custom annotation summary toolkit within the annotation tool DRAM. This toolkit, containing nearly 1300 genes for 8 different traits, enables the rapid profiling of functional potential in engineered systems, such as biofilm formation, osmoprotection, and generation of corrosive sulfide. Together, this approach provides new

insights into spatial and temporal patterns of microbial functional potential and demonstrates how -omics tools can be applied to make informed decision for reservoir management in the terrestrial subsurface.

4.2 Introduction

In 2022, the United States produced more natural gas than it consumed for the third year in a row and recovered 35.8 trillion cubic feet of natural gas – the highest annual amount ever recorded[1]. Hydraulic fracturing of subsurface shale and sandstone formations accounted for roughly 86% of this production[1]. In fact, shale gas from hydraulic fracturing is projected to remain an important component of the U.S. energy portfolio in the coming years, with an estimated ~195 billion barrels oil and ~1,700 trillion cubic feet natural gas resources remaining over an estimated >400,00mi^2 of shale basin area still available for unconventional drilling[2].

Microorganisms are ubiquitous and key players in hydraulically fractured shale systems[3–20]. During the process of hydraulic fracturing, microorganisms are introduced into the newly formed hydraulically fractured subsurface ecosystem through various inputs used[4,7,21]. Many of these microorganisms colonize and persist despite a relatively harsh environment characterized by elevated temperatures, salinity, and pressure[4]. The physiochemical conditions of the shale environment strongly affect the metabolisms and other functional traits of the persisting shale microbiome, while the emergent properties from the shale microbiomes likely also affect the overall ecosystem function. For example, microorganisms must be able to tolerate biocides, elevated salinity, and high temperatures, but may contribute to 'souring' and scaling of the well due to hydrogen sulfide production[12,22,23], bioclogging due to biofilm formation[4,24,25], and corrosion from produced sulfides and organic acids produced – all of which may decrease

hydrocarbon recovery[3,3,16,22,26]. Thus, there is a crucial need to understand functional potential and possible impacts of the microbial community in this economically important ecosystem.

Previous studies have largely focused on understanding the taxonomic composition of the shale microbiome within a single basin, or the interplay of a microbial metabolisms within a single basin[3,7,14,17,26,27]. However, fluids from shale formations can vary greatly in salinity, with some basins exhibiting brine-level salinities while others are more analogous to ocean water[28,29]. Microbial communities in highly saline fractured shales are dominated by fermentative and methanogenic microorganisms, as other metabolisms are likely constrained by energetic cost of balancing osmotic stress from high salinities[30]. However, lower saline conditions may allow for the proliferation of sulfate reducing microorganisms that can contribute to toxic hydrogen sulfide and souring of the well[31]. Many other characteristics of subsurface shales also vary greatly across geographically separated basins including temperature, rock porosity, and mineral and organic carbon content[4,32]. Given the physiochemical heterogeneity of subsurface shales and the effect this can have on the functional and taxonomic composition of the persisting microbial community, shale microbiomes must be interrogated at a large, cross-basin scale in order to make informed decisions on microbial management to reduce their possible deleterious effects.

Here, we built a comprehensive genomic database of 209 shale metagenomes spanning 34 hydraulically fractured wells across nine shale basins in North America, as well as two international basins (UK & China), to better study and inform management on the effects of persisting shale microbiomes at a basin-scale. Given the depth of genomic data, we developed an annotation toolkit (MAP-Frac) within the tool DRAM to quickly and efficiently identify and summarize genes related to key traits, such as salt tolerance, biofilm formation, etc. We leveraged the MAP-Frac toolkit with the 978 medium- and high-quality genomes and >7million

unique genes in our database to uncover differences in functional potential of shale microbiomes across geographically separated and physiochemically heterogenous shale basins. The database, toolkit, and results presented here have the potential to inform microbial management practices in hydraulically fractured shale wells, which remain an important energy resource in the United States and globally.

4.3 Results & Discussion

4.3.1 Building a comprehensive fractured shale microbiome database

To build a standardized genomic database of basin-scale shale microbiomes, we performed metagenomic sequencing on 21 produced fluid samples and combined them with 188 in house and some previously published metagenomes from the Appalachian Basin[3,17,26,31], Anadarko Basin[7], and Denver-Julesburg (DJ) Basin[20]. The resulting database of 209 metagenomes contained over 4.7 terabases of raw sequencing information (averaging 22.8 gigabase bases per sample) representing nine shale basins in North America (Appalachian, Anadarko, Denver-Julesburg, Permian, Illinois, Powder River, Michigan, Western Canadian, and Western Gulf) and two international basins: China (Sichuan Basin) and the United Kingdom (Bowland Shale) (**Figure 3.1**). All raw metagenomic reads from all 209 samples from 11 shale basins were processed through a standardized computational pipeline in order to reduce bias, error, and other discrepancies that may have an impact on the downstream analyses and results.

A vast majority of metagenomes sequenced were produced fluid samples (*n=178*), which represent the persisting subsurface fractured shale microbiome. Of the 178 produced fluid metagenomes, a majority came from 19 different wells with timeseries sampling (*n=161*). Nearly all (~80%) of the timeseries samples had paired NMR metabolomic data (**Figure 3.1**). The

remaining metagenomes (*n=31*) were 'top side' fluid samples taken during the development of the well. Many of these were samples of the fluid mixture immediately before fracturing of the well (or 'frack fluid') (*n=10*), but other samples include the original source waters used (*n=5*), recycled produced fluids (*n=6*), drill muds (*n=2*), and other water samples collected during the development of the well (*n=8*). The numerous top side samples taken pre-fracturing are independently a unique resource and provide the opportunity potentially identify key inputs that may harbor relevant, persisting microbial community members. Together, these samples represent the colonizing and persisting microbial community and allow investigations into the taxonomic, functional, and metabolomic differences of the fractured shale microbiome across geographically separated basins.

4.3.2 Curation of a genomic catalog of fractured shale microbiomes

To investigate the taxonomic and functional differences in fractured shale microbiomes at a genome-resolved basin-scale, we leveraged the 4.7 terabases of metagenomic sequencing to recover metagenome assembled genomes (MAGs). Overall, we recovered a total of 2,912 medium & high quality (>50% completion, <10% contamination) MAGs from all 209 samples using only scaffolds >5000bp in length and a standardized computational pipeline. First, MAGs were dereplicated per basin to create a basin-specific set of unique genomes. Combining all basin-specific MAG databases resulted in 1100 unique MAGs. Finally, the 1100 MAGs were dereplicated at 99% identity to create the final shale genome database of 978 medium- and high-quality unique MAGs (**Figure 3.2**). The high degree of loss of MAGs at the first dereplication step indicates the removal of duplicate MAGs recovered through time or in multiple samples

from the same basin, but the low degree of loss while dereplicating across basins highlights the uniqueness in microbial populations in shale microbiomes at a geographic scale.

Bacteria were most represented in the final MAG database (*n=880*) complimented by many archaeal genomes (*n=98*). Bacterial phyla with the highest representation were Proteobacteria, Bacteroidota, Firmicutes A, Actinobacteriota, Thermotogota, and Patescibacteria. Notable archaeal phyla represented in the MAG database were Halobacteriota, Methanobacteriota, and Nanoarchaeota (**Figure 3.2**). Overall, 52 bacterial phyla and 9 archaeal phyla were represented. Within all phyla, a high diversity of classes, families, and genus were represented – with 99 classes, 190 orders, 336 families, and 485 unique genera. Despite often being described as low-diversity ecosystems, a large portion of this taxonomic diversity was contributed exclusively by MAGs recovered from produced fluids, including 46 phyla, 82 classes, 143 orders, 250 families, and 304 unique genera. Overall, 690 of the 978 unique MAGs (70%) were recovered from produced fluid samples. Many MAGs did not classify to the family *(n=52*), genus (*n=185*), or species level (*n=636)* despite estimations of high genome completion, highlighting the level of taxonomic novelty of microbial populations within fractured shale microbiomes.

Next, we assessed how well our MAG database captured the genomic variability across shale basins by mapping metagenomic reads from all samples to the database of 978 unique MAGs. On average, a very high proportion (83%) of metagenomic reads from produced fluid samples were recruited to the MAG database, indicating that a majority of the sequenced microbiomes were represented by the recovered genomes (**Figure 3.2**). This allowed for a genome-resolved approach into the paired taxonomic and functional potential of the microbial

communities across a diversity of shale basins without missing a large proportion of unbinned sequences.

Finally, we curated a complimenting gene database to capture the genomic potential of fractured shales as thoroughly as possible. To do this, metagenomic assemblies were filtered to only scaffolds >2500bp and called genes were clustered at 97% identity. This resulted in >16 million genes identified in all metagenomic assemblies, and >7.7 million unique genes identified across basins after clustering. A majority of genes that co-clustered came from samples within the same basin, though clustering across basins did occur. For nearly all samples (203 of 209 samples), >90% of reads were recruited to the resulting database of unique genes, averaging 94% of reads but ranging from 88% to 98% reads mapped. Together, the genome- and gene-centric databases reconstructed and presented here comprehensively capture shale microbiome genetic diversity at a basin-scale.

4.3.3 *Fractured shales are not a one-microbe-fits all ecosystem*

Surprisingly, there was no taxonomic core microbiome shared across all 11 basins (defined as a MAG present in at least 75% of samples for a basin) (**Figure 3.3**). Even at a broad taxonomic view there were no phyla present across all basins. Phyla shared across a majority (>6) of basins include members of Bacteriodota, Firmicutes A, Synergistota, and Thermotogota. Produced fluids from the DJ and Anadarko basins exhibited near similar conductivities and shared nearly identical core phyla, despite wells being geographically separated by at least 450 miles (**Figure 3.3**). Interestingly, the Sichuan Basin shared all the same core phyla with the addition of Chloroflexota and two archaea phyla, Halobacteriota and Methanobacteriota, despite being located on a separate continent from the DJ and Anadarko basins. Salinity of produced

fluids are reported to be similar between the DJ and Sichuan basins which likely drives the similarities observed in their microbiomes. Only one phylum was core for the Appalachian basin (Firmicutes F) underscoring the low alpha diversity observed in this basin (**Figure 3.5**).

Even less taxa were shared across basins at a more refined taxonomic resolution. Overall, there were only 64 genus that were core to at least one shale basin. Out of these, only 14 genera were core constituents of more than one shale basin, with only two being present in >3 distinct basins: *Thermovirga* (Synergistota phylum) was core to five shale basins and *Halanaerobium* (Firmicutes F phylum) was core to four shale basins (**Figure 3.3**). Relative abundances of these genera varied across samples, wells, and basins, though overall both were relatively high abundance across all samples where they were detected. *Thermovirga* averaged 10% relative abundance across 101 samples where this genus was detected, while *Halanaerobium* averaged 19% relative abundance across 125 samples. Interestingly, there was no overlap in the basins where *Thermovirga* and *Halanerobium* were considered core, possibly suggesting an overlap in their ecological niches or other physicochemical shale conditions that constrain their geographic distribution across the same shale basins. The Anadarko basin shared the most core genera with other basins, having 11 total core genera but only two of these being distinct to the Anadarko basin. In contrast, only one of the core genera within the Bowland shale was shared with another basin (*Shewanella*, W. Gulf basin) and all remaining core genera classified within the Proteobacteria phylum. This stark difference may be driven by differences in hydraulic fracturing methods and inputs for the Bowland shale compared to many North American shales[34].

Overall, despite strong selection pressures from elevated salinity and temperature in the fractured shale environment, there is not a conserved set of microbial taxa that will colonize and persist in all shale wells. The Baas-Becking hypothesis ‘everything is everywhere’ suggests that

dispersal is not a limiting factor to microbial community assemblage[35,36], which is also supported by the abundant reuse of equipment and infrastructure across hydraulic fracturing sites likely contributing to dispersal of shale microbiomes. Persisting shale microbial communities instead vary considerably across basins and are likely strongly influenced by the salinity of the produced fluids, completing the second half of the Baas-Becking hypothesis of 'everything is everywhere, but the environment selects'. To an extent, the taxonomic profiles of persisting shale microbiomes are influenced by which microbial metabolisms and paired salt-tolerance strategy are thermodynamically favorable under different levels of salinity. For example, different salt-tolerance strategies and the cellular energy they require (i.e. the 'salt-in' vs. accumulation of compatible solutes) may explain why two fermentative microorganisms, *Halanaerobium* and *Thermovigra*, are not core to any of the same shale basins. These findings highlight the heterogeneity in taxonomic profiles of shale microbiomes at a geographic scale and underpins the importance for culture-independent and functionally informative approaches to studying shale microbiomes.

4.3.4 Taxonomic and metabolomic diversity of shale microbiomes

Shale basin microbiomes were taxonomically different by beta diversity analyses, agreeing with the lack of shared taxa and core microbiome (**Figure 3.4**). Alpha diversity also varied considerably between basins, with the Appalachian basin samples exhibiting extremely low alpha diversity while the Michigan basin, Bowland shale, and Illinois basin exhibited alpha diversities approaching soil microbiome metrics (Shannon H' ranging 0.3-4.48) (**Figure 3.5**). Previous studies from the Appalachian basin highlighted a sharp decrease in alpha diversity through time[3,4], therefore we leveraged timeseries data to interrogate this trend across basins.

Indeed, alpha diversity decreased in all five Appalachian basin wells, however alpha diversity increased in two of five Anadarko basin wells, and increased in all six DJ basin wells. Notably, diversity continued to increase in three established DJ basin wells that were over three years (>1000 days) old when sampling began. This may indicate temporal expansion of metabolic niches in shale basins exhibiting lower salinities thus allowing for taxa with more functionally diverse metabolisms to persist in fractured shales, even over longer timescales.

Metabolomes recovered from 16 wells with timeseries within the Appalachian (*n=5*), DJ (*n=6*), and Anadarko (*n=5*) basins revealed dramatic differences in the profiles of dominant metabolites across basins (**Figure 3.6**). The DJ and Anadarko basins exhibited similar metabolomic profiles, aligning with taxonomic core observations. In particular, concentrations of organic acids increased through time in most DJ and Anadarko wells, with acetate reaching nearly 15mM in several wells. In contrast, concentrations of acetate in the Appalachian basin were often <20μM. Fermentative microorganisms producing organic acids likely contribute to the high concentrations of acetate observed in the DJ and Anadarko wells in addition to other microorganisms that can persist under lower salinity conditions of produced fluids from these basins. For example, microorganisms that respire using sulfate as the terminal electron acceptor may be able to prolificate under low saline conditions such as the DJ and Anadarko Basin but are limited at higher salinities. Sulfate reducing microorganisms that do not completely oxidize organic compounds such as lactate and butyrate instead produce acetate which likely also contribute to the high concentrations observed in these basins. There was abundant evidence for acetate production in the fractured shale genome database, with 518 MAGs encoding some genomic potential for acetate production (acetate kinase) as well as 212 MAGs with potential for butyrate production (butyrate kinase) Overall, there were 54 different metabolites detected across

these 16 wells spanning three basins (total 124 samples), but only 16 were detected within at least one sample from each basin, highlighting the metabolomic variation across shale basins microbiomes.

4.3.5 Fractured shale microbiomes are taxonomically variable, but functionally conserved

Shale microbiomes are more similar to one another functionally than taxonomically (**Figure 3.4**), as taxonomic composition (from the phylum to genus level) varied amongst basins (**Figure 3.3 and 3.7**). Shale microbiomes were less dissimilar to one another when comparing beta diversity of functional composition compared to taxonomic composition (**Figure 3.4**) (taxonomic ANOSIM: $R=0.701$, $p=0.001$; functional ANOSIM: $R=0.512$, $p=0.001$). This highlights that there is likely a functional core, despite lack of a taxonomic core. Indeed, out of the >13,000 unique KEGG annotations of functional genes in the fractured shale gene database, 398 of these were present in 100% of metagenomes. A majority of these shared genes were related to transcription, translation, and nucleotide and amino acid biosynthesis machinery. However, genes for two CRISPR-associated proteins (*cas1* and *cas2*), genes related to flagella structure and function, and many genes for transporters of peptides, metals important for enzymatic activity (iron, nickel, and zinc), and phosphate were also detected across metagenomes from all basins.

Overall, shale microbiomes were dominated by inferred fermentative microorganisms (**Figure 3.8**). In many Appalachian basin samples and some DJ basin samples inferred fermentative MAGs accounted for >90% of the microbial community. On average, MAGs inferred to be fermentative accounted for >50% of the microbial community for nearly all basins. Other prominent, but less abundant, metabolisms inferred from shale MAGs were

methanogenesis and respiratory sulfate reduction (**Figure 3.8**). Methanogens and sulfate reducing microorganisms were prominent community members in most lower-salinity shale basins, such as the Permian, Anadarko, Powder River, DJ, Sichuan, Illinois and Michigan basins. However, they were not equally distributed nor dominated by the same taxonomies. Methanogens were a significant portion of the microbial community in the Illinois basin, while sulfate reducers dominated over methanogens in most samples from the Anadarko Basin. The prevalence and abundance of methanogens in the Illinois and Michigan basins may indicate the ability for microorganisms with this type of metabolism to persist in shale wells for extremely long periods of time, as these wells were over 13 years old. Overall, this highlights the conservation of key metabolisms across shale basins, despite being taxonomically different from one another. However, preferred substrates for fermentation likely differ across basins and within them which may aid in persistence of many microorganisms carrying out functionally similar metabolisms and providing different niches.

Hydrogen sulfide can have a profound effect on the fractured shale system and hydrocarbon recovery, and thus sulfate reducing bacteria are often a target for biocide treatment in oil & gas systems[37–39]. However, there are multiple pathways by which microorganisms can produce hydrogen sulfide. We observed the potential for sulfide production in many relevant (>1% relative abundance) MAGs across basins, though encoded genes differed (**Figure 3.9**). A majority of inferred sulfide-producing MAGs encoded genomic potential to reduce thiosulfate via rhodanase or thiosulfate reductase (*phsA*) genes. However, at least one MAG with the potential for dissimilatory sulfate reduction (*dsrB*) was present in most basins. Yet, across most basins many of the most abundant potential sulfide-generating MAGs were inferred to reduce thiosulfate instead of sulfate – a finding that agrees with previous fractured shale microbiome

studies[3,16,22,23,26]. An exception of this is the Anadarko and Permian basins, where canonical sulfate reducers reached similar relative abundance to many thiosulfate reducing microorganisms. This finding agreed with gene-centric analyses of sulfate and thiosulfate reducing potential in the complimentary gene database (**Figure 3.10**).

The difference between respiratory sulfate reduction and thiosulfate reduction using rhodanase is likely a reflection of salinity level selecting for different microorganisms across shale basins. Sulfate reducing microorganisms are unable to withstand high salinities, as respiration with sulfate likely does not provide sufficient energy to synthesize compatible solutes required to balance osmotic stress[30,31]. However, fermentative microorganisms in fractured shales often encoded a rhodanase genes and are more likely to withstand high salinities. Therefore, this likely explains the abundant genomic evidence for thiosulfate reduction across basins, especially those exhibiting highest salinities. This difference could also be due to substrate availability as both sulfate and thiosulfate may originate from chemical inputs[40] or minerals leaching from the rock matrix[32]. However, sulfate is more readily oxidized compared to thiosulfate and therefore may not be as abundant of a substrate in the fractured shale environment. Leaching of sulfide minerals in fractured shales may provide a continuous source of electron acceptors that sustain respiration or fermentative metabolisms for persisting shale microbiome. In turn, the continuous production of sulfide likely contributes to production challenges such as precipitation of sulfide minerals, hydrocarbon souring, and corrosion[41].

4.3.6 The MAP-Frac toolkit identifies key microbial traits in MAGs and metagenomes

In addition to pathways for energy generation, other functional traits can play an important role in the ability for microorganisms to be transferred large distance and across

basins, as well as the ability to colonize and persist in the newly formed fractured shale environment. Microorganisms that persist are impacted by the physiochemical conditions of fractured shale ecosystems and must encode genomic potential to tolerate high salinities, temperatures, and biocide applications. However, microorganisms that do persist likely also encode for other traits that may have a direct impact on the ecosystem and hydrocarbon recovery, such as biofilm formation (that may contribute to bioclogging) and hydrocarbon degradation.

To assess these the potential for these traits within our large genome and gene database, we developed an annotation summary toolkit MAP-Frac (Microbes Affecting Production in FRACturing systems) within the annotation tool DRAM[42] to screen for relevant functional traits quickly and efficiently (**Figure 3.11**) The MAP-Frac toolkit identifies and summarizes genes related to eight traits of interest in fractured shale ecosystems: biocide resistance (integration of the BACMET database[43]), sporulation, biofilm formation, salt tolerance strategies, sulfide generation, hydrocarbon degradation (integration of the CANT-HYD database[44]), heat tolerance, and production of organic acids. In total, the MAP-Frac toolkit summarizes annotations of 566 genes related to these eight traits, as well as includes a blast database of 743 genes and 54 HMMs. Applying this toolkit to our MAG database revealed differences in the functional potential for these traits in microbiomes across basins.

Salt tolerance strategies varied across basins amongst the dominant MAGs. For example, many of the most abundant MAGs in the high-saline Appalachian basin encoded multiple copies of genes related to the salt-in strategy as well as genes related to uptake and synthesis of compatible solutes, such as glycine betaine, ectoine, trehalose, etc. (**Figure 3.12**). In contrast, MAGs that dominate lower salinity basins such as the Illinois and Michigan basin encoded fewer genes related to all strategies. Microorganisms that use the salt-in strategy under high saline

conditions must encode a specialized proteome that can withstand high internal salt concentrations and are highly adapted to high salinity environments. This likely limits their distribution across lower saline shale basins, where instead microorganisms are inferred to use compatible solutes as osmoprotectants, and may in part explain the lack of a taxonomic core microbiome. Importantly, many osmoprotectants such as choline, glycine betaine, trehalose and others have been shown to be important substrates for fermentative microorganisms as well – thereby connecting a key trait of osmoprotection to the dominant energy generating metabolism that together likely contribute significantly to microbial persistence in fractured shale ecosystems.

Interestingly, the median number of genes related to heat tolerance that a MAG encoded was very similar across all shale basins (average 11 genes per MAG) despite shale formation temperatures ranging in temperature (generally 50-125°C)[28]. The most commonly encoded heat tolerance genes were two chaperon proteins (*dnaK*, *dnaJ)*, and HSP20 (small heat shock protein 20). Notably, these proteins are also been observed in cells under osmotic, pH, heavy metal, and antibiotic stress in addition to thermal stress[45,46]. This suggests that they likely play a role in stabilizing host proteins and machinery in fractured shales due to a variety of harsh conditions besides elevated temperature and may aid in biocide resistance. Microbial populations that encode more or highly express genes for chaperon proteins therefore may have a competitive advantage to persisting in the fractured shale environment.

Anaerobic hydrocarbon degradation genes were identified in MAGs through an HMM screen via integration of the CANT-HYD database[44] into the FRAC-MAP toolkit. This revealed limited genomic potential for the anaerobic degradation of hydrocarbons despite inferred abundance of them in the fractured shale envrioment. Only ten MAGs encoded a putative alkane

C2 methylene hydroxylase (*ahyA*) indicating potential for anaerobic alkane degradation. Four of these MAGs belonged to the genus *Sulfurospirillum A*, while others were members of the families Aeromonadaceae and Steroidobacteraceae and two novel genera within the phylum Myxococcota. Nearly half of all genomes (~45%) with the potential for aerobic or anaerobic degradation (*n=19* different degradation genes detected) were recovered from topside samples. This has important implications for the storage and treatment of produced fluids the mitigate the risks of microorganisms degrading target hydrocarbons in both the subsurface and topside aspects of hydraulic fracturing systems.

Finally, sporulation genes may be of critical importance to understanding the spread of pervasive microorganisms in fractured shale ecosystems. Much of the drilling and fracturing machinery and infrastructure is reused across hydraulic fracturing sites, and thus may act to seed subsurface shale microbiomes. However, microorganisms that spread must be able to live anaerobically in the subsurface as well as tolerate oxic conditions during transfer and spread. The number of genes and type of genes related to sporulation varied greatly across MAGs present in different basins (**Figure 3.13**), and therefore may help to identify targets for antimicrobial treatments of hydraulic fracturing machinery.

4.4 Conclusions

Here, we constructed a genomic database of shale microbiomes from 209 metagenomes spanning a diversity of shale basins – mainly across North America. Overall, we leveraged 978 unique medium- and high-quality MAGs for a genome-resolved approach into fractured shale microbiomes. By first taxonomically profiling shale microbiomes across basins, we observed the lack of a core microbiome even at the phylum level. Less taxa were shared across all shale

basins, with genera *Thermovirga* and *Halanaerobium* being the most prevalent. Despite of the lack of core taxa, we observed the conservation of key metabolisms such as fermentation, sulfate and thiosulfate reduction, and methanogenesis. However, functional potential and specific pathways within these metabolisms varied across MAGs and basins. This underscores the importance for constructing a genomic database of fractured shale microbiomes and leveraging a genome-centric approach to studying fractured shale microbiomes. Additionally, we built a custom annotation summary toolkit to efficiently screen genomes and genes for key traits of interest in our genomic catalog. These traits, such as salt tolerance, heat tolerance, biocide resistance, and sporulation, likely play an important role in microbial dispersal, colonization, and persistence across fractured shale ecosystems. Overall, we leveraged the genome database to profile metabolisms and key traits of shale microbiomes to understand how the functional potential of persisting microorganisms differ across basins, and how they may impact the overall shale ecosystem and hydrocarbon recovery. Therefore, the results presented here have the potential to inform microbial management practices across an important energy system.

4.5 Materials & Methods

4.5.1 Sample collection, DNA extraction, and metagenomic sequencing

Samples were collected from many collaborators and coauthors who have previously worked in fractured shale systems. Many samples were previously published as individual datasets where methods of sample collection, extraction, and sequencing are detailed, including the Appalachian basin[3,26,31], Anadarko basin[7], and Denver-Julesburg basin[20]. Permian and Western Gulf basin samples were processed according to methods in Amundson *et al*. 2022. All additional samples were sent to CSU for quantification via the High Sensitivity Qubit Assay and

sent to University of Colorado, Anschutz Medical Campus for metagenomic sequencing according to methods described in Amundson *et al*., 2022. For the Illinois and Michgian basin samples (contributed by Dr. Anna Martini, Amherst College and Dr. Julian Damasheck, Utica University), an unknown amount of produced fluids were filtered and DNA was extracted using MoBio UltraClean Soil DNA Isolation Kit. For Western Canadian Basin samples (contributed by Dr. Casey Hubert and Gabrielle Scheffer, University of Calgary), 200mL samples of produced fluids were filtered through a 0.2-micron filter and DNA was extracted via Qiagen PowerSoil DNA extraction kit. Sample processing for the Bowland shale, Powder River basin, and Sichuan Basin are unknown.

4.5.2 Building a fractured shale genomic catalog via metagenomic assembly, annotation, and binning

All raw metagenomic reads were processed identically through the same in house bioinformatic pipeline regardless of if the dataset had been previously published or not. In some cases, this resulted in re-processing of some previously published samples from the Appalachian Basin using updated tools. Samples from some basins did not assemble well and multiple methods (individual assembly and co-assemblies) were leveraged to increase the quality of the metagenomic assembly and/or number of genomes recovered from the sample. First, raw metagenomic reads were first trimmed with Sickle (https://github.com/najoshi/sickle) from 5' to 3' ends and assembled with IDBA-UD using default parameters[43]. Some samples (*n=17*) were also assembled with MEGAHIT (v.1.2.9)[44] to improve gene and genome recovery from these metagenomes.

Resulting scaffolds >5000bp from all metagenomic assemblies (*n=226*) were used for subsequent binning of bacterial and archaeal genomes using MetaBAT2 (v2.12.1)[45], resulting in metagenome assembled genomes (MAGs). Quality and completeness of genomes were evaluated using the lineage workflow of CheckM (v.1.1.2) ('lineage_wf') followed by the 'qa' command[46]. Only resulting MAGs meeting medium (>50% completion and <10% contamination) and high (>90% completeness and <5% contamination) quality standards[47] were carried through to subsequent steps. For each basin, MAGs were first dereplicated at a strain-level of 99% ANI (default parameters) for each basin using dRep (v.3.0.0)[48]. Basin-unique MAGs were then combined into one database of 1100 MAGs which were then dereplicated using dRep to 978 unique medium- and high-quality MAGs resulting in the fractured shale genome database. Taxonomic classifications of these genomes were determined using GTDB-tk (v.2.1.1)[49]. Finally, MAGs were annotated with DRAM (v.1.4.4) to uncover functional potential[38].

To build the gene database for fractured shales, all metagenomic assemblies (*n=226*) were first filtered for only scaffolds >2500bp in length. Genes (nucleotide sequences) on scaffolds were called using prodigal (v.2.6.3)[50] and clustered at 97% identity using MMseqs2[51] easy-linclust and alignment mode 3. Finally, amino acid sequences for the resulting gene database were annotated via DRAM (v.1.4.4)[38].

4.5.3 Chemical and metabolite analyses

Conductivity was measured on all raw produced fluid samples from the Appalachian, Anadarko, and Denver-Julesburg basin using a Myron L 6PIIFCE meter. Metabolomics was also performed on all samples for these three basins at the Pacific Northwest National Laboratory in the Environmental Molecular Sciences Laboratory. Methods for the Appalachian basin

metabolomics were previous published in Daly *et al*., 2016[3]. Samples from the Anadarko and DJ basins were processed according to methods published in Amundson *et al*., 2022[7].

4.5.4 Determining relative abundance of genomes and genes in the fractured shale database

Relative abundance of MAGs and genes in the fractured shale database was determined by mapping trimmed metagenomic reads. However, metagenomes were first rarified to standardize for size due to a large variation in sequencing depth. Metagenomic reads from metagenomes >10Gbp were first rarified to 10Gbp to decrease the range from 3-99Gbp to 3-10Gbp using the 'reformat' guide within bbmap[52]. Metagenomes were not rarified to the smallest metagenome size of 3Gbp as this would dramatically reduce the amount of sequencing data.

Coverage of genomes was determined by multi-mapping trimmed and rarified metagenomic reads to MAGs using bbmap (v.38.89)[52] with minID=0.90. The sam files generated were convered to bam files using samtools (v.1.9)[53]. Bam files were sorted and filtered for minID=0.95 using the reformat guide in bbmap. Individual coverages for each MAG was determined with coverM (https://github.com/wwood/CoverM) (v0.6.0) using filtered and sorted bam files. Two commands were run to determine coverage. First, coverM was run using –min-covered-fraction=90 to determine MAGs read recruitment to at least 90% of the genome. Second, coverage values were calculated using the -m reads_per_base command, which represents reads mapped/genome length, and thus multiplied this by read length (151bp) in order to calculate MAG coverage (simply, coverage = reads_per_base * 151 bp). Final coverage values were obtained from trimmed mean values normalizing coverage values ('-m trimmed_mean').

Coverage of genes was determined using the same tools as coverage calculations for MAGs. However, a higher identity (minID=0.97) was used for filtering resulting bam files, and

coverM was used in ‘contig mode’ as opposed to ‘genome mode’. For both genomes and genes, relative abundance was calculated as the proportion of coverage for a gene or MAG out of the total coverage for all genes or MAGs within sample.

4.5.5 Taxonomic profiling of shale microbiomes

In the absence of 16S rRNA amplicon sequencing data for all samples, metagenomes were instead taxonomically profiled using singleM (https://github.com/wwood/singlem), which uses strings of 20 short amino acid sequences within single copy genes to profile taxonomic composition via metagenomic reads. Rarified metagenomic reads were used to standardize across sequencing depths. This was executed by feeding rarified metagenomic reads to the ‘pipe’ function with flages ‘--otu-table’ to output the entire OTU table and ‘--taxonomic-profile’ to output a condensed taxanomic profile output.

4.5.6 Core microbiome analyses

The taxonomic core microbiome for each shale basin and across shale basins was determined using singleM profiles of metagenomes using >1x coverage for singleM in determining profiles. An OTU of specific taxonomic classification was required to be present in at least 75% of produced fluid samples, regardless of relative abundance, in order to be considered core to a specific basin. The lack of a core microbiome at the phylum level across basins was due to the absence of any basin-core phyla being core across all basins.

4.5.7 Build the FRAC-Map toolkit

The FRAC-MAP toolkit is built within the annotation tool DRAM (v.1.4.4)[38] and is executable within the 'distill' step using the –custom_distillate flag. This toolkit summarizes genes related to salt tolerance strategies, biofilm formation, sporulation, biocide resistance, heat tolerance, hydrocarbon degradation, sulfide generation, and production of organic acids. Hydrocarbon degradation is identified via Hidden Markov Models (HMMs) of hydrocarbon degradation genes within the CANT-HYD database[40] using the confidence scores defined in the publication of this resource. Biocide resistance genes are identified from known and putative genes in the BACTMET database[39] using the default BLAST bit scores within DRAM. Genes for salt tolerance strategies were previous described in Amundson *et. al*., 2022.

4.5.8 Diversity, multivariate and statistical analyses

Beta diversity (Bray-Curtis) was calculated in R (v.4.2.3) using the package vegan (2.6-4) for taxonomic diversity using singleM profiles from rarified metagenomic reads. For functional diversity, only genes annotated via KEGG were considered for beta diversity (Bray-Curtis) metrics. Non-metric multidimensional scaling ordination (NMDS) ordinations display this diversity using phyloseq. Alpha diversity (Shannons') was also determined from singleM profiles. Between group differences within NMDS ordinations based on beta diversity were determined from Analysis of Group Similarities (ANOSIM) calculations. Principal component analysis (PCA) of metabolite data was also done in vegan (v.2.6-4) and significant vectors (<0.05) were determined form envfit within vegan in R.

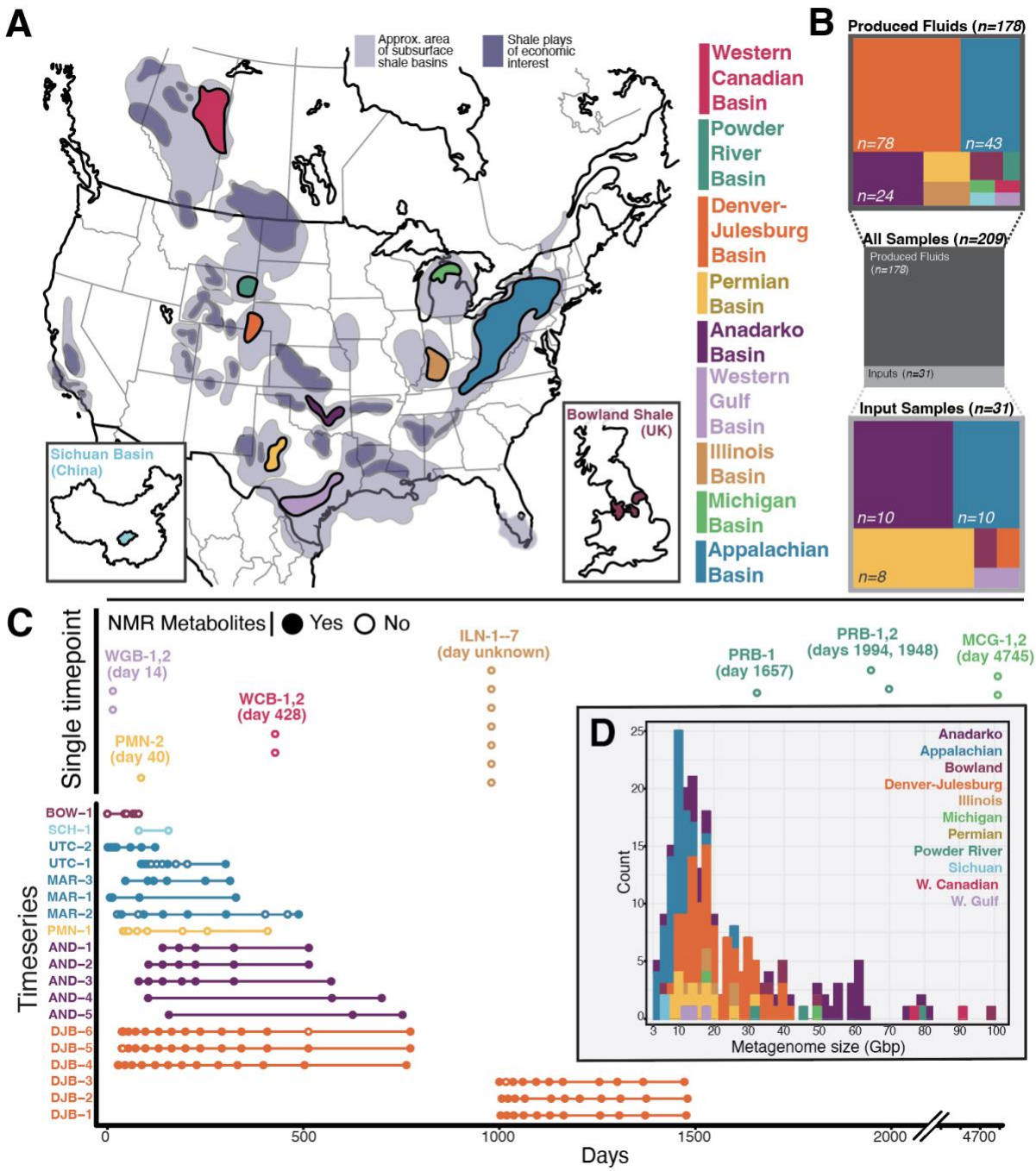

Figure 3.1. Description of shale database samples. **(A)** Map of approximate shale basin locations across North America, China, and the United Kingdom. Shale basins that are represented in this dataset are colored, with their corresponding label to the left of the map. **(B)** Treemaps depicting the total samples in the shale database, broken down by number of produced fluid vs. input samples and colored by the number of type samples for a given basin. **(C)** Produced fluid samples and whether they belong to a well with timeseries sampling, or just a single timepoint. Solid circles indicate paired NMR metabolomic data for the sample. **(D)** Histogram of non-rarified metagenome sizes, in gigabase pairs (Gbp), colored by basin.

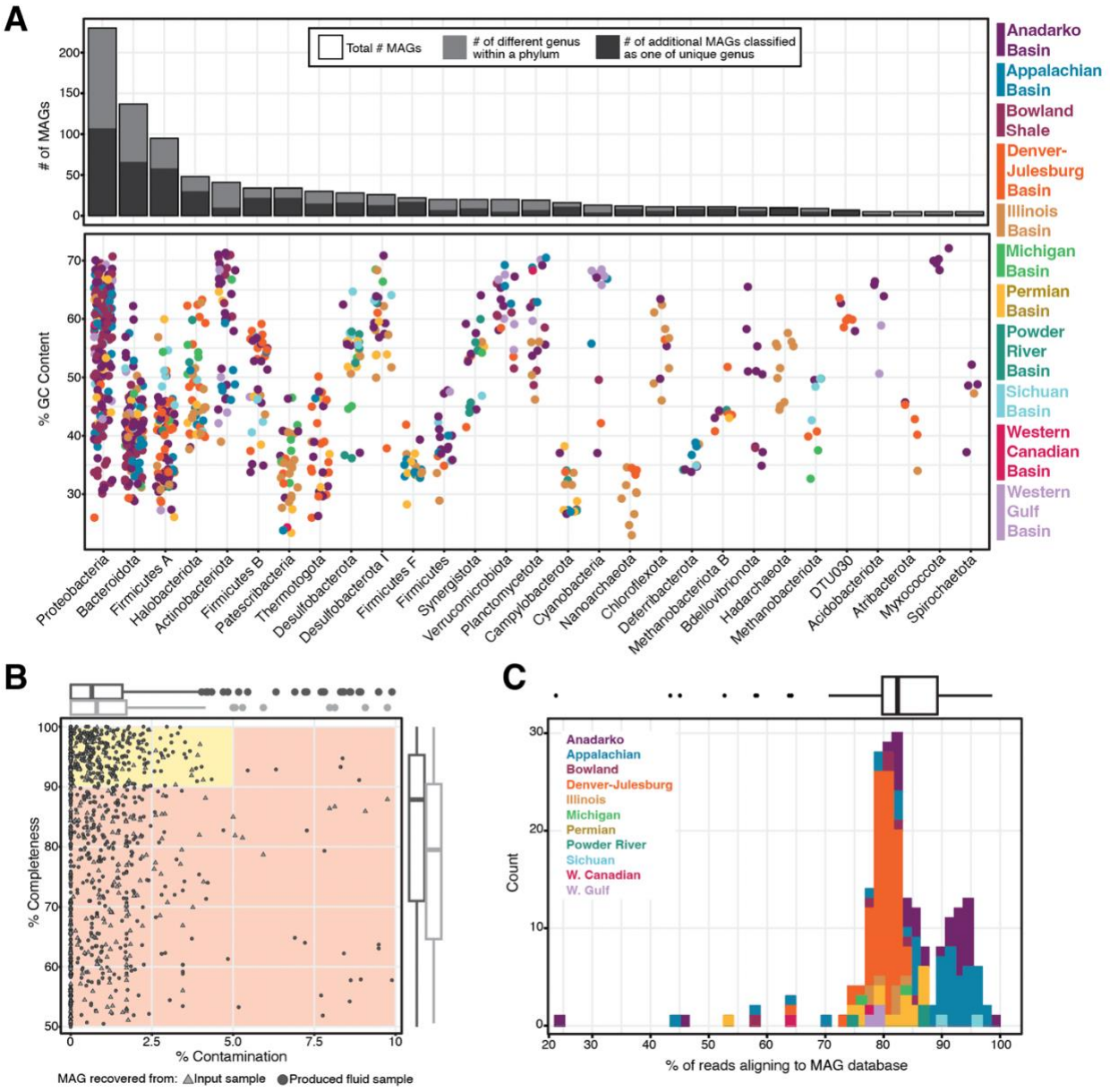

Figure 3.2. Database of 978 unique MAGs from fractured shale microbiomes. **(A)** Bottom scatter plots show the top 30 phyla with the most MAGs recovered, where each point represents a single MAG plotted by its %G-C content and colored by the basin it was recovered from. Top bar charts illustrate the total number MAGs within a phyla, and how many different genus are represented within a given phyla (light grey bars). **(B)** Estimates of contamination and completeness for all MAGs. Background shading indicates classifications of medium quality (orange) and high quality (yellow). Boxplots outside the plots depict the distribution of completeness (right margin) and contamination (top margin) across all MAGs, with medians shown as thick lines. MAGs and boxplots are split by the type of sample the MAG was recovered from, either input samples (triangles, light grey) or produced fluid samples (circles, dark grey). **(C)** Histogram and boxplot showing distribution of the percentage of trimmed and rarified metagenomic reads that aligned to the database of 978unique MAGs, colored by basin.

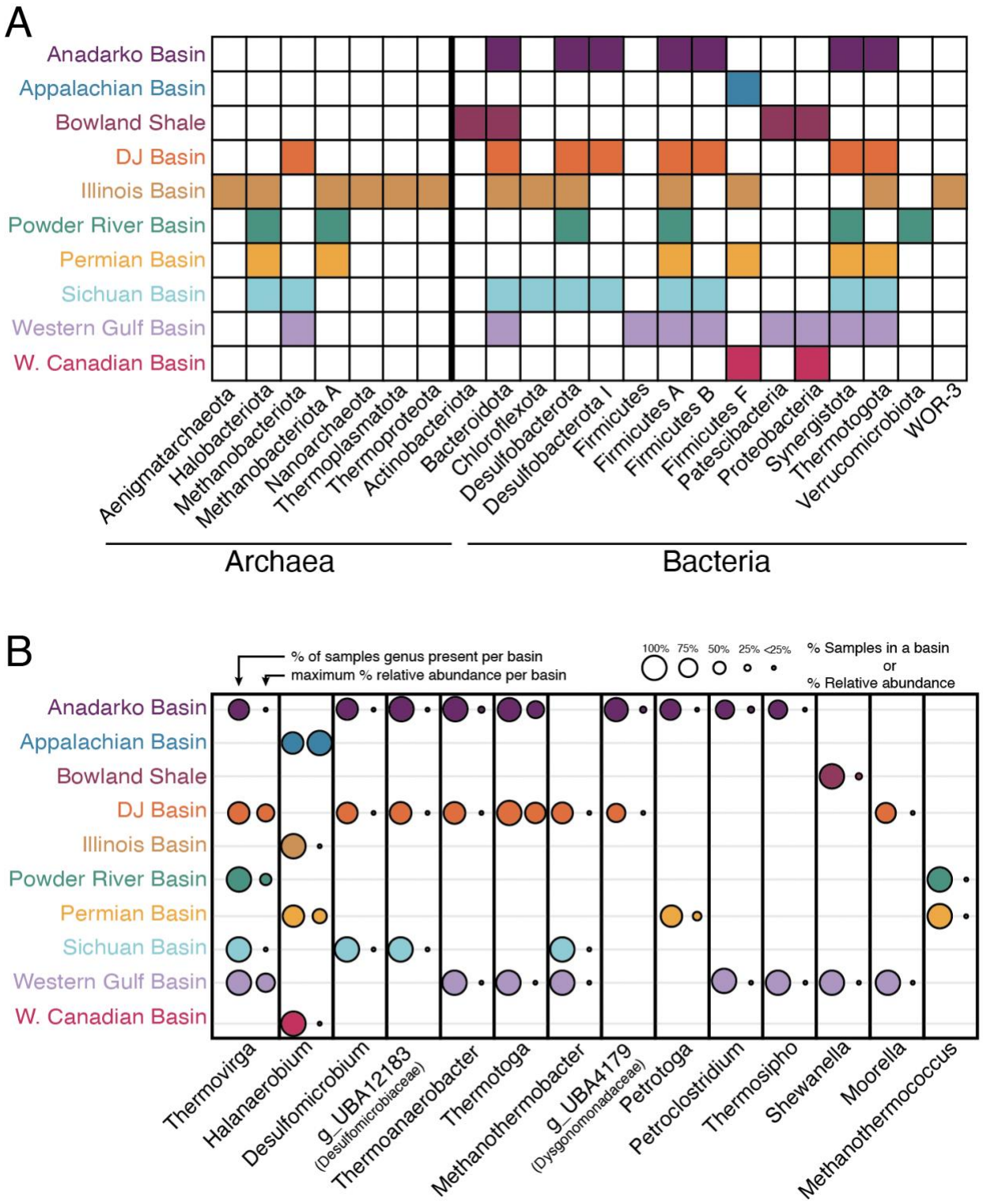

Figure 3.3. Core microbiomes profiles. For any taxonomic level (phylum or genus), taxa had to be present in >75% of samples to be considered core. **(A)** Phyla that are core to each basin. **(B)** Genera core to each basin. Two circles are present for each genus and are sized by: the percentage of samples that a given genus was present in for a given basin (left column), the maximum percent relative abundance that a genus achieved in a given basin (right column).

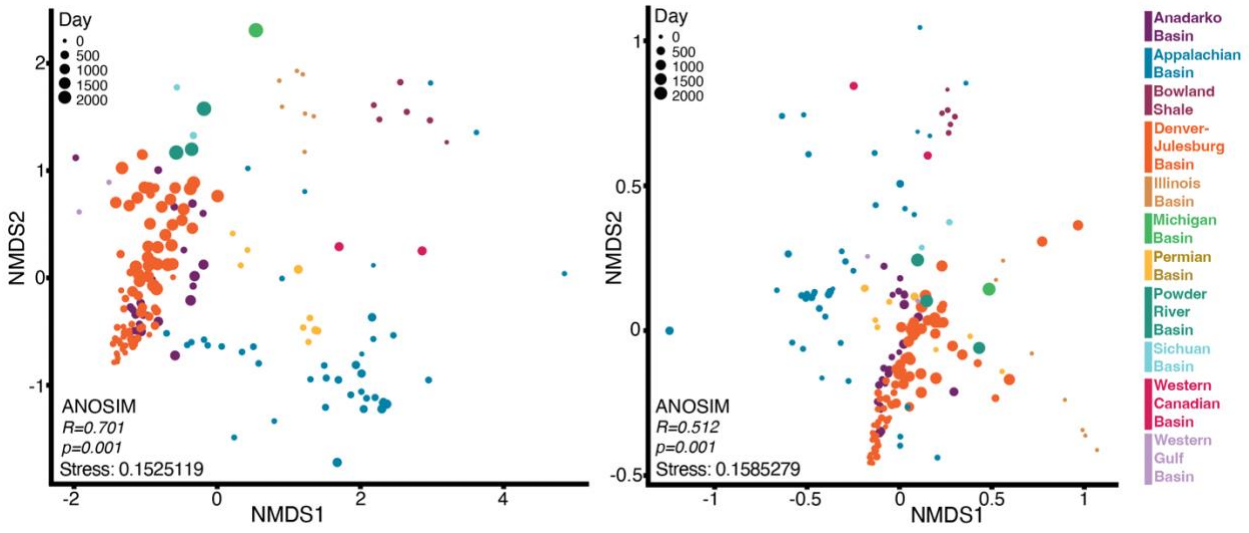

Figure 3.4. Non-metric multidimensional scaling (NMDS) ordinations of Bray-Curtis beta diversity of taxonomic composition (left) and functional composition (right), colored by basin. Taxonomic composition was determined by profiling metagenomes using singleM and determining relative abundance. Functional composition was determined by relative abundance of genes across samples, as calculated from recruitment of metagenomic reads to the fractured shale gene database.

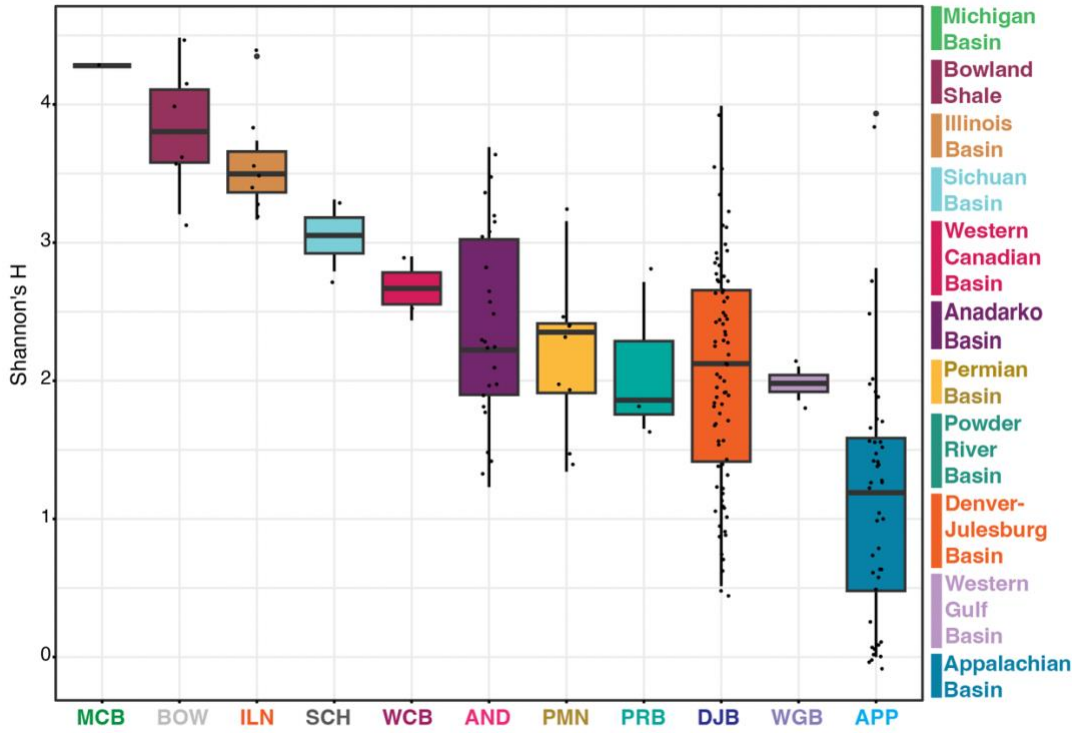

Figure 3.5. Alpha diversity (Shannons' H) by basin, organized by decreasing alpha diversity. Basin abbreviations at bottom (left to right) match full names (top down on right).

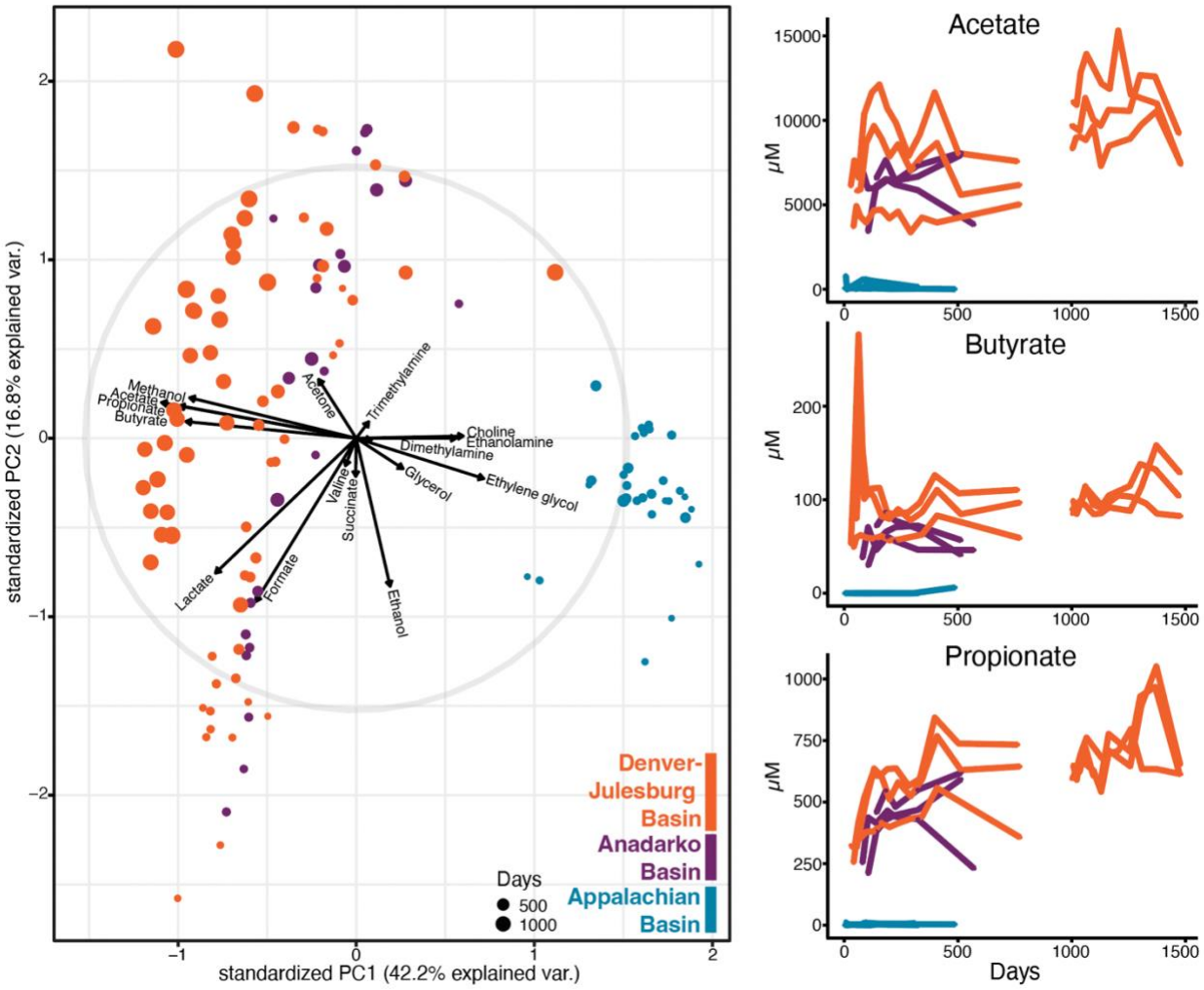

Figure 3.6. (Left) Principal component analysis (PCA) of NMR metabolite data from three shale basins: Anadarko, DJ, and Appalachian. Each point represents a single sample's metabolomic composition, colored by basin and sized by day. Vectors show metabolites that significantly (<0.05) explain differences between metabolomes. **(Right)** Temporal measurements of some organic acids (acetate, butyrate, and propionate) for the three basins.

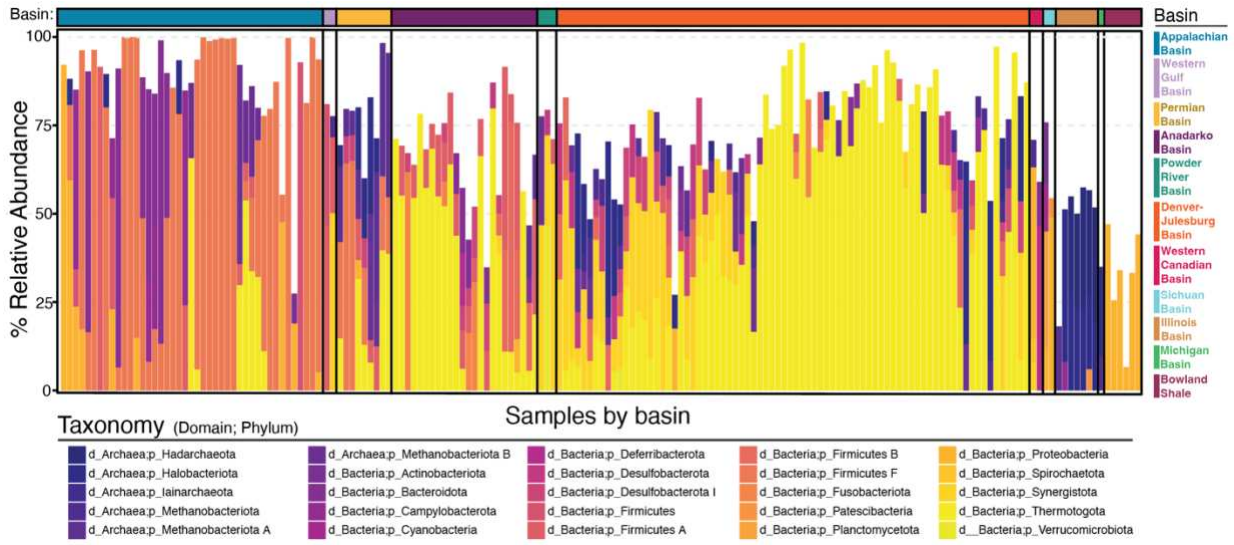

Figure 3.7. Barcharts showing the taxonomic profiles of dominant (>5% relative abundance) microorganisms across shale basins. Remaining space between bar chart and 100% represent taxa <5% relative abundance. Each bar represents a single sample, organized by days and basin (top horizontal bar) and colored by phylum.

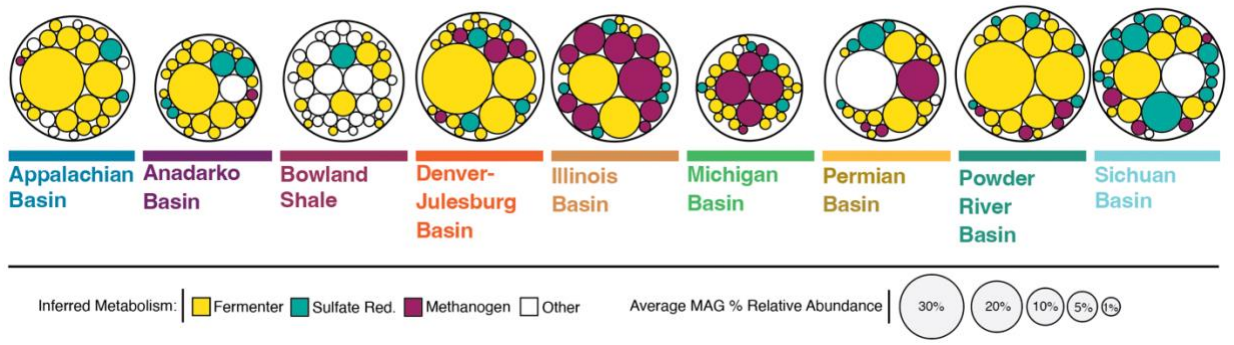

Figure 3.8. Bubble plots depicting the dominance of inferred fermentative MAGs (yellow), inferred methanogens (purple), and inferred sulfate reducers (teal). MAGs with other possible metabolisms are left white. MAGs are depicted as circles, sized by their average relative abundance per basin.

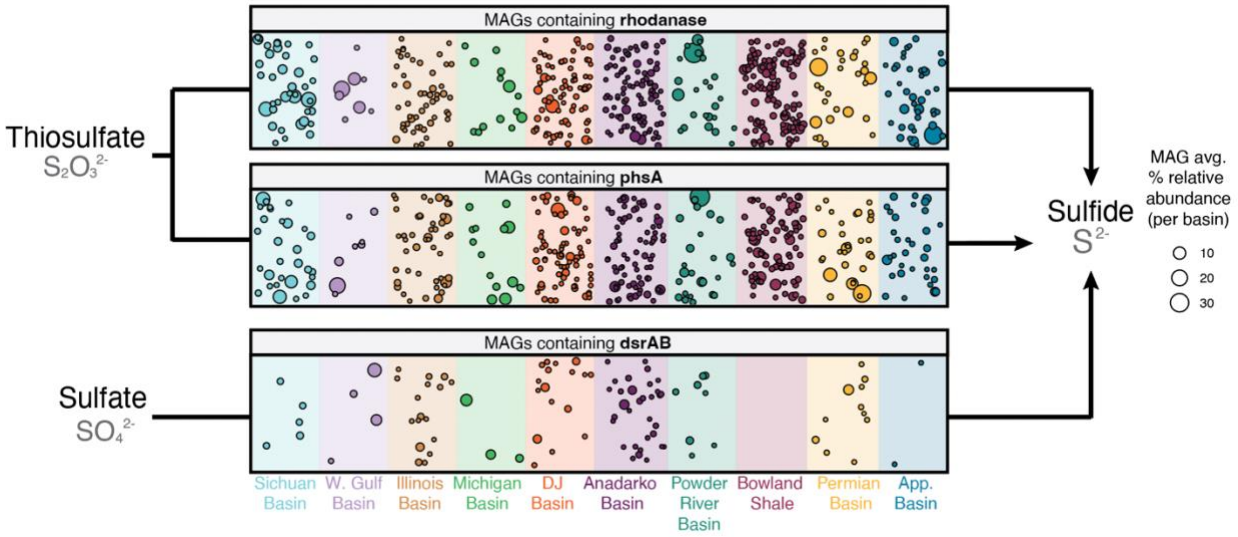

Figure 3.9. Genomic potential in MAGs for production of hydrogen sulfide through three pathways, shown by identifying marker genes for these pathways in MAGs. Shaded background columns illustrate basins, while each circle represents an individual MAG, sized by its average relative abundance per basin. MAGs (circles) are arranged by functional potential (rows) and basin (columns).

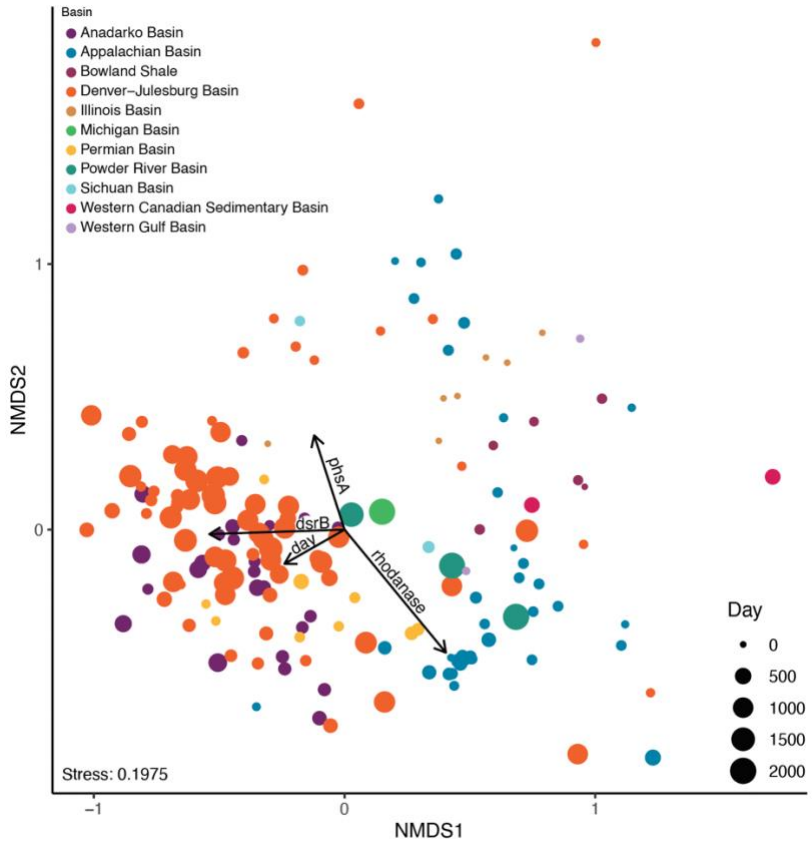

Figure 3.10. Non-metric multidimensional scaling (NMDS) ordination of Bray-Curtis beta diversity (calculated from gene relative abundances) of sulfur cycling genes across metagenomes. Points are colored by basin and sized by day. Vectors illustrate significantly correlated (<0.05) variables that explain variation in the distribution of datapoints.

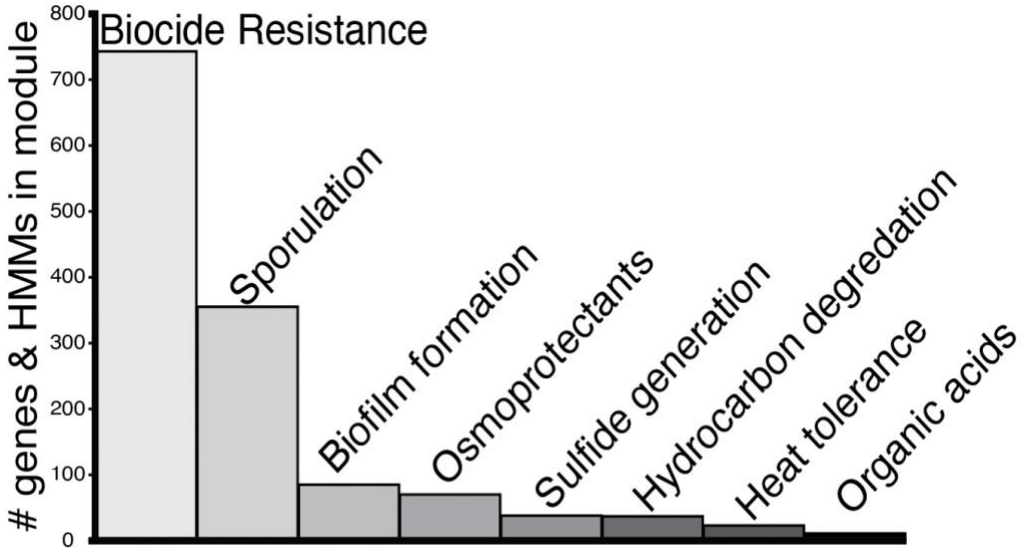

Figure 3.11. Number of genes and/or HMMs for each trait within the FRAC-MAP summary toolkit executed within the annotation tool DRAM.

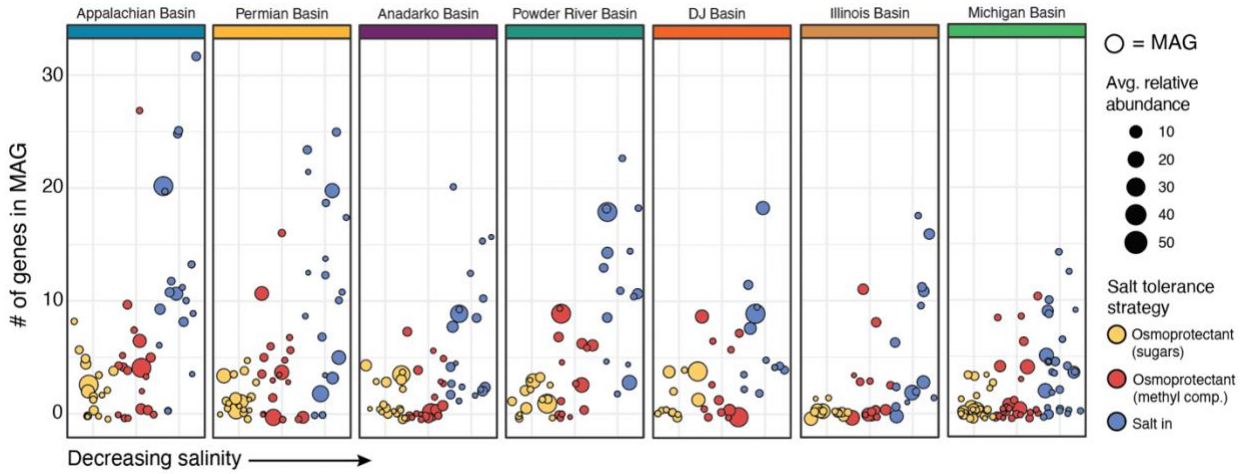

Figure 3.12. Salt tolerance strategies for dominant (>1% relative abundance) MAGs across a subset of shale basins. Each circle represents an individual MAG, sized by its average relative abundance for a given basin. Basins are ordered approximately by decreasing salinity. MAGs are colorde by their genomic potential for different salt tolerance strategies (osmoprotectants: yellow and red, salt-in strategy: blue).

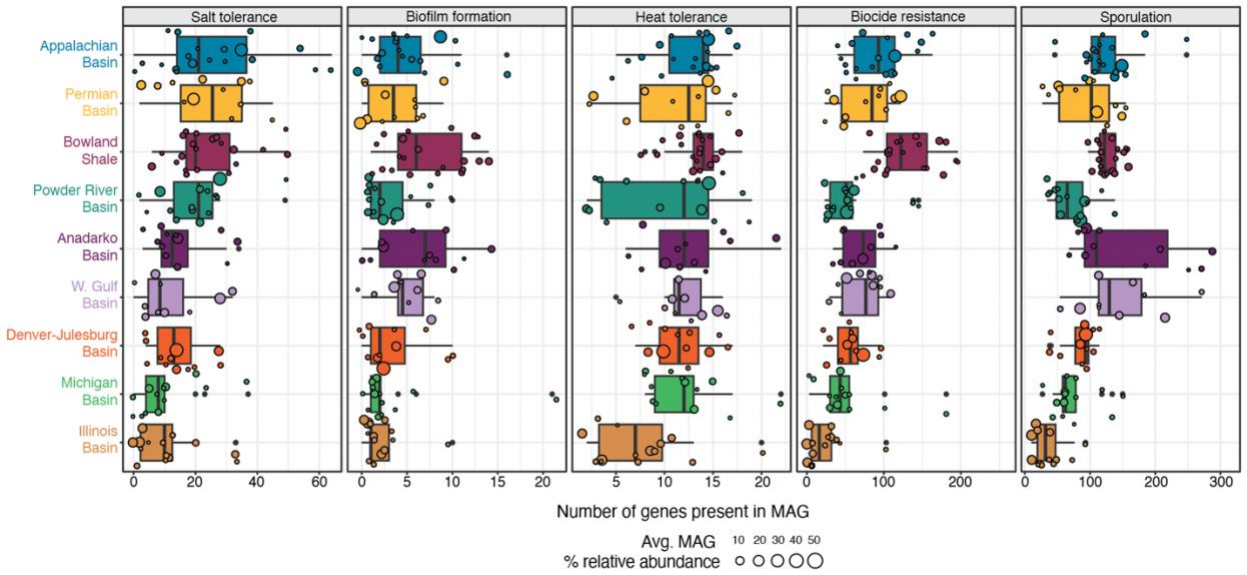

Figure 3.13. Summary of number of genes encoded in MAGs for key traits across shale basins. Shale basins are approximately organized from highest to lowest salinity.

Chapter 4 References

1. Where our natural gas comes from - U.S. Energy Information Administration (EIA). https://www.eia.gov/energyexplained/natural-gas/where-our-natural-gas-comes-from.php.

2.Assumptions to the Annual Energy Outlook 2023: Oil and Gas Supply Module. (2023).

3. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).

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

5. Struchtemeyer, C. G. & Elshahed, M. S. Bacterial communities associated with hydraulic fracturing fluids in thermogenic natural gas wells in North Central Texas, USA. *FEMS Microbiol. Ecol.* **81**, 13–25 (2012).

6. Bacterial Communities Associated with Production Facilities of Two Newly Drilled Thermogenic Natural Gas Wells in the Barnett Shale (Texas, USA) | SpringerLink. https://link.springer.com/article/10.1007/s00248-012-0073-3.

7. Amundson, K. K. *et al.* Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. *Microbiome* **10**, 5 (2022).

8. Stemple, B. *et al.* Biogeochemistry of the Antrim Shale Natural Gas Reservoir. *ACS Earth Space Chem.* **5**, 1752–1761 (2021).

9. Schlegel, M. E., McIntosh, J. C., Bates, B. L., Kirk, M. F. & Martini, A. M. Comparison of fluid geochemistry and microbiology of multiple organic-rich reservoirs in the Illinois Basin, USA: Evidence for controls on methanogenesis and microbial transport. *Geochim. Cosmochim. Acta* **75**, 1903–1919 (2011).

10. Zhong, C. *et al.* Comparison of the Hydraulic Fracturing Water Cycle in China and North America: A Critical Review. *Environ. Sci. Technol.* **55**, 7167–7185 (2021).

11. Zhong, C. *et al.* Temporal Changes in Microbial Community Composition and Geochemistry in Flowback and Produced Water from the Duvernay Formation. *ACS Earth Space Chem.* **3**, 1047–1057 (2019).

12. An, B. A., Shen, Y. & Voordouw, G. Control of Sulfide Production in High Salinity Bakken Shale Oil Reservoirs by Halophilic Bacteria Reducing Nitrate to Nitrite. *Front. Microbiol.* **8**, (2017).
13. Wang, H., Lu, L., Chen, X., Bian, Y. & Ren, Z. J. Geochemical and microbial characterizations of flowback and produced water in three shale oil and gas plays in the central and western United States. *Water Res.* **164**, 114942 (2019).
14. Evans, M. V. *et al.* Members of Marinobacter and Arcobacter Influence System Biogeochemistry During Early Production of Hydraulically Fractured Natural Gas Wells in the Appalachian Basin. *Front. Microbiol.* **9**, (2018).
15. Murali Mohan, A. *et al.* Microbial Community Changes in Hydraulic Fracturing Fluids and Produced Water from Shale Gas Extraction. *Environ. Sci. Technol.* **47**, 13141–13150 (2013).
16. Lipus, D. *et al.* Predominance and Metabolic Potential of Halanaerobium spp. in Produced Water from Hydraulically Fractured Marcellus Shale Wells. *Appl. Environ. Microbiol.* **83**, e02659-16 (2017).
17. Borton, M. A. *et al.* Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale. *Environ. Microbiol.* **20**, 4596–4611 (2018).
18. Cluff, M. A., Hartsock, A., MacRae, J. D., Carter, K. & Mouser, P. J. Temporal Changes in Microbial Ecology and Geochemistry in Produced Water from Hydraulically Fractured Marcellus Shale Gas Wells. *Environ. Sci. Technol.* **48**, 6508–6517 (2014).
19. Mohan, A. M., Bibby, K. J., Lipus, D., Hammack, R. W. & Gregory, K. B. The Functional Potential of Microbial Communities in Hydraulic Fracturing Source Water and Produced Water from Natural Gas Extraction Characterized by Metagenomic Sequencing. *PLOS ONE* **9**, e107682 (2014).
20. Amundson, K. K., Roux, S., Shelton, J. L. & Wilkins, M. J. Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales. *Curr. Biol.* **33**, 3125-3135.e4 (2023).
21. Struchtemeyer, C. G., Davis, J. P. & Elshahed, M. S. Influence of the Drilling Mud Formulation Process on the Bacterial Communities in Thermogenic Natural Gas Wells of the Barnett Shale. *Appl. Environ. Microbiol.* **77**, 4744–4753 (2011).

22. Booker, A. E. *et al.* Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* **2**, 10.1128/mspheredirect.00257-17 (2017).
23. Liang, R. *et al.* Metabolic Capability of a Predominant Halanaerobium sp. in Hydraulically Fractured Gas Wells and Its Implication in Pipeline Corrosion. *Front. Microbiol.* **7**, (2016).
24. Booker, A. E. *et al.* Deep-Subsurface Pressure Stimulates Metabolic Plasticity in Shale-Colonizing Halanaerobium spp. *Appl. Environ. Microbiol.* **85**, e00018-19 (2019).
25. Vikram, A., Lipus, D. & Bibby, K. Metatranscriptome analysis of active microbial communities in produced water samples from the Marcellus Shale. *Microb. Ecol.* **72**, 571–581 (2016).
26. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci.* **115**, E6585–E6594 (2018).
27. Nixon, S. L. *et al.* Genome-Resolved Metagenomics Extends the Environmental Distribution of the Verrucomicrobia Phylum to the Deep Terrestrial Subsurface. *mSphere* **4**, 10.1128/msphere.00613-19 (2019).
28. Shaffer, D. L. *et al.* Desalination and Reuse of High-Salinity Shale Gas Produced Water: Drivers, Technologies, and Future Directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
29. Cooper, C. M. *et al.* Oil and Gas Produced Water Reuse: Opportunities, Treatment Needs, and Challenges. *ACS EST Eng.* **2**, 347–366 (2022).
30. Oren, A. Thermodynamic limits to microbial life at high salt concentrations. *Environ. Microbiol.* **13**, 1908–1923 (2011).
31. Oren, A. Life at High Salt Concentrations. in *The Prokaryotes: Prokaryotic Communities and Ecophysiology* (eds. Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 421–440 (Springer, 2013). doi:10.1007/978-3-642-30123-0_57.
32. Khan, H. J. *et al.* A Critical Review of the Physicochemical Impacts of Water Chemistry on Shale in Hydraulic Fracturing Systems. *Environ. Sci. Technol.* **55**, 1377–1394 (2021).
33. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).

34. Hernandez-Becerra, N. *et al.* New microbiological insights from the Bowland shale highlight heterogeneity of the hydraulically fractured shale microbiome. *Environ. Microbiome* **18**, 14 (2023).
35. Green, J. L., Bohannan, B. J. M. & Whitaker, R. J. Microbial Biogeography: From Taxonomy to Traits. *Science* **320**, 1039–1043 (2008).
36. Hanson, C. A. Microbial Biogeography. in *International Encyclopedia of Geography* 1–6 (John Wiley & Sons, Ltd, 2017). doi:10.1002/9781118786352.wbieg0231.
37. Johnson, K., French, K., Fichter, J. K. & Oden, R. Use Of Microbiocides In Barnett Shale Gas Well Fracturing Fluids To Control Bacteria Related Problems. in (OnePetro, 2008).
38. Struchtemeyer, C. G., Morrison, M. D. & Elshahed, M. S. A critical assessment of the efficacy of biocides used during the hydraulic fracturing process in shale natural gas wells. *Int. Biodeterior. Biodegrad.* **71**, 15–21 (2012).
39. Bhagobaty, R. K. Culture dependent methods for enumeration of sulphate reducing bacteria (SRB) in the Oil and Gas industry. *Rev. Environ. Sci. Biotechnol.* **13**, 11–16 (2014).
40. Elsner, M. & Hoelzer, K. Quantitative Survey and Structural Classification of Hydraulic Fracturing Chemicals Reported in Unconventional Gas Production. *Environ. Sci. Technol.* **50**, 3290–3314 (2016).
41. Gieg, L. M., Jack, T. R. & Foght, J. M. Biological souring and mitigation in oil reservoirs. *Appl. Microbiol. Biotechnol.* **92**, 263–282 (2011).
42. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
43. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. & Larsson, D. G. J. BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* **42**, D737–D743 (2014).
44. Khot, V. *et al.* CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Front. Microbiol.* **12**, (2022).
45. Gophna, U. & Ron, E. Z. Virulence and the heat shock response. *Int. J. Med. Microbiol.* **292**, 453–461 (2003).

46. Fay, A. *et al.* The DnaK Chaperone System Buffers the Fitness Cost of Antibiotic Resistance Mutations in Mycobacteria. *mBio* **12**, 10.1128/mbio.00123-21 (2021).
47. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
48. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/31/10/1674/177884.
49. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
50. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
51. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
52. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
53. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
54. Prodigal: prokaryotic gene recognition and translation initiation site identification | BMC Bioinformatics | Full Text. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-119.
55. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets | Nature Biotechnology. https://www.nature.com/articles/nbt.3988.
56. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166 (2014).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Chapter 5: Conclusion

5.1 Summary

The collective aim of this dissertation was to piece together the key traits and metabolisms that aid in microbial colonization and persistence in subsurface fractured shales. Specifically, I aimed to understand how and why these traits and metabolisms vary across heterogenous shale basins, and how that variability may impact the overall system and potential hydrocarbon recovery. Here I achieved this by using a multi-omic approach to study the functional potential and metabolomes of shale microbiomes across many shale basins that differed in a suite of physiochemical conditions.

I first applied this metagenomic approach to characterize the functional potential of a shale microbiome in a basin that exhibited very different salinities and temperatures compared to most other previously published fractured shale microbiome studies (Chapter 2)[1]. Through this, I provide evidence for prevalent respiratory sulfate reducing microorganisms in addition to thiosulfate reducing microorganisms in a single shale basin. Lower saline fluids from this basin likely allowed for an expanded metabolic potential of the persisting microorganisms[2,3] – a trend that was also observed in Chapter 4 when considering many geographically distinct shale basins. In addition, within Chapter 2 I provide a potential framework describing how degradation of complex polymers commonly used in the hydraulic fracturing process may be a key trait that provides carbon sources and electron donors for other microbial community members, thereby aiding in persistence of shale microbiomes.

In Chapter 3, I focus on viral predation and how host defense against viruses may be a crucial trait for persisting microorganisms[4]. A large proportion of genomes recovered from datasets in Chapters 2 and 3 encoded genes for CRISPR-cas viral defense, as well as other viral

defense systems. Viruses are the only predator in fractured shales[5] and as such host defense may be necessary for survival as host and viral populations continue to interact in this closed ecosystem. However, despite abundant evidence for CRISPR viral defense and a possible shift toward arrays that may be more effective, I observed an increase in host-virus coexistence through time. This highlights the proliferation of viruses in face of host defense systems and potentially hints at the limits of CRISPR as a key trait protecting against viral predation.

Finally, I leveraged a large amount of metagenomic sequencing to study the biogeographical patterns of some of these same traits and metabolisms (Chapter 4) using insights and data obtained from Chapters 2 and 3. In this study I built a comprehensive gene and genome catalog and associated annotation toolkit to uncover patterns in the distribution of metabolisms (namely fermentation, methanogenesis, and respiratory sulfate reduction) and other key traits across geographically distinct shale basins. I expand on the results from Chapters 2 and 3 to consider other traits that may contribute the persistence of shale microbiomes, such as hydrocarbon degradation, biocide resistance, heat tolerance, and sporulation. The physical heterogeneity of shale basins likely has a strong impact on the microorganisms that are able to colonize and persist. Specifically, basins with lower saline produced fluids harbor increased functional diversity as the energetic cost of counteracting high salinity decreases and other microbial metabolisms become favorable. However, understanding what these functional differences are within persisting shale microbiomes is key to understanding the impacts the shale microbiomes may have in turn on the broader system.

Overall, the body of research presented in this dissertation provides a holistic view of shale microbiomes by studying functional potential through metagenomics. The results presented in Chapters 2, 3, and 4 provide new insights into the functional traits that contribute to microbial

colonization and persistence in fractured shales and have the potential to inform microbial management practices in hydraulic fracturing systems.

5.2 Beyond fractured shales: implications for other subsurface ecosystems

Insights gained from fractured shale wells could be applied to many other engineered subsurface ecosystems. Microbial life has been reported across many conventional oil, gas, and coalbed systems and is likely influenced by many of the same physiochemical conditions as fractured shales (i.e., elevated temperature and salinity, highly reduced redox conditions, etc.)[6,7]. Additionally, most produced fluid from fracking wells is currently disposed of in subsurface wells that have been depleted of their resources and have been converted to 'storage wells'[8]. These shale wells, as well as other subsurface cavities, also have been a target for other industrial waste, as well as carbon capture and storage projects for mitigating the effects of climate change by preventing CO_2 from reentering the atmosphere[9]. Finally, there is rapidly growing interest in drilling for hydrogen as an energy source[10] or for subsurface storage[11] – both engineered subsurface system with direct implications to U.S. energy resources that may benefit from insights derived from fractured shale microbial communities.

The subsurface is also the largest reservoir for life on earth[12–14]. Though fractured shale ecosystems are not pristine subsurface ecosystems, they do offer a view into microbial life in the subsurface[15]. Microorganisms that survive in fractured shales likely exhibit many of the same traits and generate energy through some of the same metabolisms as microorganisms persisting in other pristine subsurface ecosystems. However, microbes in pristine subsurface ecosystems tend to grow very slowly as these systems are often very space limited and extremely oligotrophic[16]. Microorganisms in fractured shales, on the other hand, have access to more

connected physical space as well as nutrients (as provided by leaching minerals from the shale or additives used in the development of the well)[15]. Therefore, insights derived from research on microbial colonization and persistence in fractured shales can be applied to not only many other engineered subsurface ecosystems, but also potentially provide insights into microbial life in pristine ecosystems within the deep biosphere.

5.3 Implications for produced water reuse

Hydraulic fracturing uses millions of gallons of freshwater in order to develop and fracture an individual shale well. Though exact numbers vary by state and region, hydraulic fracturing uses approximately two to seven million gallons of fresh water per well, and potentially up to 30 million gallons in some regions such as shale plays underlying Oklahoma[17]. This fresh water has been transformed to saline water when it returns to the surface in the form of flowback and produced fluids. Not all injected water returns to the surface, with estimates of returning fluid generally ranging between 30-70%[17] of total injected water, but being reported as low as 8%[18]. This results in an estimate of nearly 25 billion gallons of produced water being generated every year[19], resulting in two main challenges: (1) water resources may be further strained in regions already under water scarcity which likely impact the surrounding communities and (2) massive amounts of produced water must be treated, re-injected as storage or for reuse in hydraulic fracturing, or otherwise managed. Therefore, through hydraulic fracturing we are depleting freshwater resources and producing large amounts of saline water with complex chemical compositions that must be further studied.

Currently, a majority of produced fluid is reinjected into the subsurface for permanent storage[19]. However, reuse of produced fluids for hydraulic fracturing does occur but varies by

region, with highest reuse and recycling reported for produced fluids from the Marcellus formation in Pennsylvania[19]. However, two major challenges in reusing produced water for hydraulic fracturing of new shale wells are the residual microbial community and elevated starting salinities of produced fluids[20]. The microbial communities that persist in fractured shales and return to the surface have already been selected for as microorganism that can withstand physiochemical conditions of fractured shale ecosystems. Therefore, reusing this water is of concern, as it may potentially seed new shale wells with a microbiome that is adapted to the downhole conditions and may potentially cause production challenges, such as sulfidogenesis and bioclogging, to arise more quickly. However, understanding the functional potential of the persisting microbial community could help to inform better treatment of the produced water prior to reuse and in turn help reduce the chances that these production challenges will appear more quickly compared to if fresh water was used.

Additionally, a growing body of literature is investigating the potential use of produced water for irrigation of crops[21–25]. However, produced fluids often require pre-treatment or large dilutions to decrease salinity and de-toxify other heavy metals or naturally occurring radioactive compounds that may have the potential to bioaccumulate or otherwise be released into the environment. Though there are different treatments to desalinate produced fluids, many are not feasible at the scale of water produced from hydraulic fracturing[26].

In both cases, reusing produced fluids or applying them as irrigation, understanding the microbial community of fractured shales is crucial. High salinity is arguably the biggest challenge to using produced fluids for irrigation and therefore produced fluids from shale basins exhibiting lower salinities may be more promising for this application. However, as shown here, there is increased diversity in functional potential of the shale microbiome under lower saline

conditions which could have implications for produced water chemistry or impacts on the soil microbiome when applied. Therefore, understanding what the functional potential of persisting microbiomes across shales with varying salinities could help to inform better treatment and reuse of produced water in the United States and elsewhere.

5.4 Future research directions

5.4.1 Supporting metagenomic data with activity measurements

A clear gap in the body of literature focused on studying fractured shale microbiomes is the absence of microbial activity measurements through metatranscriptiomics or metaproteomics in most studies. These techniques, measuring the activity of microbiomes through RNA transcripts or protein composition respectively, provide needed insights to help validate the genomic potential gained from metagenomic data. Metatranscriptomics and metaproteomics also offer a view of the *in situ* microbiome function which potentially provides an advantage over laboratory experiments or incubations using produced fluids. Few studies have leveraged these approaches[27–30] and there exists a need to holistically evaluate the activity of fractured shale microbiomes from field samples in addition to laboratory experiments. For example, some of the most low abundance microorganisms have been observed to be very active transcriptionally, and analyzing gene copy numbers in genomes recovered through metagenomics is not always reflective of the overall *in situ* expression. Field samples are often challenging to process quickly in order to preserve RNA and proteins within a cell and thus remains a barrier to using these activity measuring techniques. However, more laboratory experiments leveraging these techniques by testing specific hypotheses generated from metagenomic data are necessary to gain

confident understanding of microbiome activity and how this may sustain microbial life in fractured shales or impact the overall ecosystem.

5.4.2 Investigations into other relevant microbial metabolisms

A majority of studies, including the work here, has highlighted fermentation, methanogenesis, and sulfate reduction as the most prevalent and commonly reported microbial metabolisms across in fractured shale ecosystems. However, under lower salinities more dissimilatory processes could become thermodynamically favorable as the energetic cost to balance osmotic stress is reduced. More research is needed into other possible relevant metabolisms, specifically leveraging the dataset presented in Chapter 4 which spans includes shale basins that of vary in salinity. This idea is particularly relevant for methanogenesis pathways as lower saline conditions may allow for hydrogenotrophic methanogenesis in addition to methylotrophic methanogenesis. Understanding this is crucial, as hydrogenotrophic methanogenesis may compete for H_2 that could also be used as an electron donor for respiratory sulfate reducing microorganisms that also may be present in lower salinity basins. The prominence of other potential metabolisms, especially those related to cycling of metals (i.e. iron), is also directly influenced by substrate availability as chemical complexity of produced fluids and the fractured shale environment can lead to binding and precipitation of these minerals, thus still limiting potential for those metabolisms.

5.4.3 Enhancing biogenic methane production

Stimulating methanogens for increased production of biogenic methane has been an area of focus in coalbed systems for several years[31–34]. This concept could be applied to hydraulically

fractured shales, especially those with less community complexity due to higher salinities. Non-trivial amounts of biogenic gas have been reported from wells in the Illinois and shallow Michigan basins[35,36], potentially suggesting that that biogenic methane may be feasible to enhance the total hydrocarbon recovery from a single well. The biggest challenge to achieving this *in situ* is stimulating methanogens by using substrates that will not cause competition between methanogens and other community members. In fractured shale systems, most methanogens are methylotrophic using methylated amines or sulfides to produce methane. Thus, potentially providing these types of labile substrates to directly stimulate growth of methanogens or indirectly targeting methanogens by using amendments that fermentative microorganisms may use to produce methanogenic substrates (as was highlighted in the Appalachian basin[37]) may be a possible way to enhance methane production in fractured shales.

5.4.4 Applying machine learning to predict shale microbiome function

A substantial amount of metagenomic sequencing exists as a result of this dissertation research in addition to sequencing data deposited from other previously published studies. This sequencing data could be leveraged collectively for machine learning applications to possibly predict the taxonomic or functional potential of shale microbiomes. However, there are barriers to taking this approach that must be overcome.

First, there are significant computational challenges associated with compiling, processing, and analyzing large amount of metagenomic sequencing data. High-capacity computational resources (external servers) are needed to handle this type of data, in addition to training resources for those conducting this research. The annotation summary toolkit presented in Chapter 4 may contribute to a solution for part of this challenge by streamlining analyses that

help to uncover key functional traits of shale microbiomes. Another barrier is the historical lack of standards for metadata and sequencing data deposition. Sequencing data is sometimes poorly or inaccurately described on public repositories, and/or necessary metadata may be missing from the associated publications. This makes mining published sequencing data to leverage with machine learning a challenge. Still, applying machine learning to predict fractured shale microbiomes is an exciting possibility and should be further investigated, especially as machine learning methods progress and become more widespread across microbiome datasets[38,39]. Finally, timeseries data is likely a requirement for predictions with machine learning. In order to train a model reliably, samples must be taken from the same well over time to understand how the microbial community composition (functional or taxonomic) progresses. The dataset presented in Chapter 4 offers the opportunity to apply these techniques given the amount of timeseries data for nineteen different wells.

In applying machine learning methods, data must be split into training and testing datasets. Thus, even though metagenomic data is rich in genomic content, sample size is often the limiting factor in applying these types of approaches. One potential solution to this challenge is using a genome-resolved approach through metagenomics and by leveraging MAGs as observations and their inferred functional potentials and relative abundance through time and over different shale wells as classification variables. This would greatly increase the number of observations and potentially allow for better training by predicting whether or not a MAG of a specific function will be present or not in a given basin.

If successful, machine learning offers the opportunity to predict the persisting shale microbiome and has the potential to substantially improve microbial management in fractured shales. In an ideal scenario, a trained model may be used to predict the persisting microbial

community at later timepoints (for example, >100 days) using few samples from early days post-fracturing (for example <100 days). Information gleaned from the approach could inform management by proactively targeting microorganisms that are predicted to increase in abundance and may negatively impact hydrocarbon recovery. For example, if successful, this approach could predict the portion of the community that has potential for generation of toxic hydrogen sulfide which contributes to souring of a well. Over time, prediction capabilities would improve as additional samples are incorporated into the model and used for training data. This would reduce the need for large amounts of metagenomic sequencing and alleviate some of the computational strains associated with understanding microbiome function yet help to predict what possible deleterious impacts the microbial community may have on the fractured shale system. Thus, machine learning this is promising application in combination with metagenomic data to inform microbial management in these systems and applications of this should be explored in the future.

5.4.5 Metagenome informed microbial management

Eradicating microorganisms entirely from hydraulically fractured shale systems is highly unlikely. In order to do so, massive amounts of time, money, and energy would need to be invested into sterilization of the infrastructure as well as the necessary inputs used such as source waters or drill muds that are crucial in developing a shale well. Even if this level of sterilization was feasible for all hydraulic fracturing wells, it is still unlikely that all microorganisms would be removed from the system and could even select for those that are able to withstand biocide treatments. Instead, studying the persisting microbiome of fractured shales may help to inform better management practices that may reduce the deleterious effects by microorganisms. For

example, more dedicated studies of inputs and machinery could help to pinpoint the 'hotspot' sources of dominant, persisting microorganisms which may even identify machinery that facilitates microbial dispersal across these systems at large geographic scales. As discussed, machine learning may have perhaps the most potential for informing management practices by proactively treating for microorganisms that are predicted to be abundant and potentially damaging at later timepoints before they become more abundant. However, further studies on successful treatment, especially under subsurface conditions, is crucial for successful applying the results presented here at scale.

In all, the research presented in this dissertation builds on literature describing taxonomic profiles of microbial communities in hydraulically fractured shales by using metagenomics to uncover the functional potentials – which could ultimately be leveraged for better microbial management in these economically important ecosystems.

Chapter 5 References

1. Amundson, K. K. *et al.* Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. *Microbiome* **10**, 5 (2022).
2. Oren, A. Thermodynamic limits to microbial life at high salt concentrations. *Environ. Microbiol.* **13**, 1908–1923 (2011).
3. Oren, A. Life at High Salt Concentrations. in *The Prokaryotes: Prokaryotic Communities and Ecophysiology* (eds. Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F.) 421–440 (Springer, 2013). doi:10.1007/978-3-642-30123-0_57.
4. Amundson, K. K., Roux, S., Shelton, J. L. & Wilkins, M. J. Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales. *Curr. Biol.* **33**, 3125-3135.e4 (2023).
5. Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
6. Nikolova, C. & Gutierrez, T. Use of Microorganisms in the Recovery of Oil From Recalcitrant Oil Reservoirs: Current State of Knowledge, Technological Advances and Future Perspectives. *Front. Microbiol.* **10**, 2996 (2020).
7. Pannekens, M., Kroll, L., Müller, H., Mbow, F. T. & Meckenstock, R. U. Oil reservoirs, an exceptional habitat for microorganisms. *New Biotechnol.* **49**, 1–9 (2019).
8. Clark, J. E., Bonura, D. K. & Van Voorhees, R. F. An Overview of Injection Well History in the United States of America. in *Developments in Water Science* (eds. Tsang, C.-F. & Apps, J. A.) vol. 52 3–12 (Elsevier, 2005).
9. Matter, J. M. & Kelemen, P. B. Permanent storage of carbon dioxide in geological reservoirs by mineral carbonation. *Nat. Geosci.* **2**, 837–841 (2009).
10. Kovač, A., Paranos, M. & Marciuš, D. Hydrogen in energy transition: A review. *Int. J. Hydrog. Energy* **46**, 10016–10035 (2021).
11. Zhang, F., Zhao, P., Niu, M. & Maddy, J. The survey of key technologies in hydrogen energy storage. *Int. J. Hydrog. Energy* **41**, 14535–14552 (2016).
12. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6578–6583 (1998).
13. Gold, T. The deep, hot biosphere. *Proc Natl Acad Sci USA* (1992).

14. Flemming, H.-C. & Wuertz, S. Bacteria and archaea on Earth and their abundance in biofilms. *Nat. Rev. Microbiol.* **17**, 247–260 (2019).
15. Mouser, P. J., Borton, M., Darrah, T. H., Hartsock, A. & Wrighton, K. C. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS Microbiol. Ecol.* **92**, fiw166 (2016).
16. Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* **11**, 83–94 (2013).
17. Shaffer, D. L. *et al.* Desalination and Reuse of High-Salinity Shale Gas Produced Water: Drivers, Technologies, and Future Directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
18. Generation, transport, and disposal of wastewater associated with Marcellus Shale gas development - Lutz - 2013 - Water Resources Research - Wiley Online Library. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/wrcr.20096.
19. Cooper, C. M. *et al.* Oil and Gas Produced Water Reuse: Opportunities, Treatment Needs, and Challenges. *ACS EST Eng.* **2**, 347–366 (2022).
20. Khan, H. J. *et al.* A Critical Review of the Physicochemical Impacts of Water Chemistry on Shale in Hydraulic Fracturing Systems. *Environ. Sci. Technol.* **55**, 1377–1394 (2021).
21. Yang, Y. *et al.* Safety and Technical Feasibility of Sustainable Reuse of Shale Gas Flowback and Produced Water after Advanced Treatment Aimed at Wheat Irrigation. *ACS Sustain. Chem. Eng.* **10**, 12540–12551 (2022).
22. Echchelh, A., Hess, T. & Sakrabani, R. Reusing oil and gas produced water for irrigation of food crops in drylands. *Agric. Water Manag.* **206**, 124–134 (2018).
23. Pica, N. E., Carlson, K., Steiner, J. J. & Waskom, R. Produced water reuse for irrigation of non-food biofuel crops: Effects on switchgrass and rapeseed germination, physiology and biomass yield. *Ind. Crops Prod.* **100**, 65–76 (2017).
24. Kondash, A. J. *et al.* The impact of using low-saline oilfield produced water for irrigation on water and soil quality in California. *Sci. Total Environ.* **733**, 139392 (2020).
25. Miller, H. *et al.* Reusing oil and gas produced water for agricultural irrigation: Effects on soil health and the soil microbiome. *Sci. Total Environ.* **722**, 137888 (2020).

26. Do Thi, H. T., Pasztor, T., Fozer, D., Manenti, F. & Toth, A. J. Comparison of Desalination Technologies Using Renewable Energy Sources with Life Cycle, PESTLE, and Multi-Criteria Decision Analyses. *Water* **13**, 3023 (2021).
27. Vikram, A., Lipus, D. & Bibby, K. Metatranscriptome analysis of active microbial communities in produced water samples from the Marcellus Shale. *Microb. Ecol.* **72**, 571–581 (2016).
28. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci.* **115**, E6585–E6594 (2018).
29. Booker, A. E. *et al.* Deep-Subsurface Pressure Stimulates Metabolic Plasticity in Shale-Colonizing Halanaerobium spp. *Appl. Environ. Microbiol.* **85**, e00018-19 (2019).
30. Booker, A. E. *et al.* Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* **2**, 10.1128/mspheredirect.00257-17 (2017).
31. Liu, B. *et al.* Improved formation of biogenic methane by cultivable bacteria in highly volatile bituminous coals. *J. Clean. Prod.* **366**, 132900 (2022).
32. Jones, E. J. P., Voytek, M. A., Corum, M. D. & Orem, W. H. Stimulation of Methane Generation from Nonproductive Coal by Addition of Nutrients or a Microbial Consortium. *Appl. Environ. Microbiol.* **76**, 7013–7022 (2010).
33. Chen, R., Bao, Y. & Zhang, Y. A Review of Biogenic Coalbed Methane Experimental Studies in China. *Microorganisms* **11**, 304 (2023).
34. Barnhart, E. P. *et al.* In Situ Enhancement and Isotopic Labeling of Biogenic Coalbed Methane. *Environ. Sci. Technol.* **56**, 3225–3233 (2022).
35. Martini, A. M. *et al.* Genetic and temporal relations between formation waters and biogenic methane: Upper Devonian Antrim Shale, Michigan Basin, USA. *Geochim. Cosmochim. Acta* **62**, 1699–1720 (1998).
36. Martini, A. M., Walter, L. M. & McIntosh, J. C. Identification of microbial and thermogenic gas components from Upper Devonian black shale cores, Illinois and Michigan basins. *AAPG Bull.* **92**, 327–339 (2008).
37. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).

38. Ghannam, R. B. & Techtmann, S. M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Comput. Struct. Biotechnol. J.* **19**, 1092–1107 (2021).
39. Namkung, J. Machine learning methods for microbiome studies. *J. Microbiol.* **58**, 206–216 (2020).

Appendix A: Chapter 2 Supplementary Discussion

Osmoprotection is a key physiological trait for microorganisms persisting in deep subsurface shale ecosystems

Deep shales frequently contain highly saline fluids that likely derive from the dissolution of salt minerals[2–4]. Under these conditions, microorganisms that persist within the fractured shale network must expend energy combatting osmotic stress while still maintaining enough energy for critical cellular processes[5]. This can be achieved through salt-in strategies, which were first observed and characterized in strict halophiles[6,7], as well as biosynthesis or uptake of compatible solutes, such as sugars, amino acids, betaine, and ectoine[8]. We observed that osmoprotection strategies are still a necessary physiological trait of microorganisms in the STACK formation, despite produced fluids exhibiting lower salinities relative to other formations, such as the Appalachian Basin or Bakken formation (**Figure 2.9**).

In general, many of the MAGs recovered in this study encoded both compatible solute and 'salt-in' mechanisms to combat salinity stress (**Figure 2.9**). Genes encoding 'salt-in' strategies were broadly distributed across many STACK MAGs, mainly via Na+:H+ antiporters (*mnhABCDEFG* and *nhaC*) and potassium uptake (*trkAH* and *ktrAB*). This may suggest that the persisting microorganisms may respond to high salinity via a bi-phasic response of initial import of cations, such as K^+, and subsequent uptake or synthesis of compatible solutes, such as proline or glutamate[9–11]. Ultimately, the rapid intake of cations allows microorganisms to quickly adjust and combat osmotic stress, while high concentrations of charged compatible solutes balances intracellular cations and stabilizes cellular structures in the long term. While genomic potential for compatible sugars was less notable, some MAGs still contained potential for synthesis of trehalose, proline, and mannitol, as well as genes related uptake of other diverse sugar

compatible solutes. Many MAGs also contained genomic potential for the uptake of ectoine (*ehuABCD*), which is one of the most common compatible solutes within the domain Bacteria[12,13]. Ectoine has also been implicated as an key compatible solute under heat stress, which may be important to persisting MAGs in the STACK formation under high temperature conditions[14–16]. However, the source of ectoine in this system is unclear; no MAGs contained any genes for ectoine synthesis (*ectABCD*), while metabolite analysis was unable to identify significant ectoine concentrations in the external environment. Finally, the uptake of methylamine-related compounds, such as glutamine, carnitine, choline and glycine betaine, was constrained mostly to the inferred-respiratory MAGs. However, only two MAGs showed evidence of glycine betaine synthesis, both via choline (*betAB*) despite the presence of elevated choline concentrations in some HF inputs. This is in contrast to the Appalachian Basin, where synthesis of glycine betaine was shown to be a key process in supporting the persisting microbial community[17].

Limited genomic potential for quaternary amine cycling in the STACK shale play

We observed limited genomic potential for quaternary amine cycling in the STACK formation in direct contrast to previous shale studies[1,17]. The transformation of choline, and associated quaternary amines has previously been shown to support key community members in shales with brine level salinities[1,17]. This metabolism is almost completely absent in the less-saline STACK formation, despite the presence of choline in the frack fluids of both sets of wells, especially STACK-14 (2,823 μL – 3,128 μL). In fact, elevated choline concentrations in STACK-14 were a significant discriminant factor when comparing chemical profiles across the wells (**Table 2.8**). However, only four inferred respiratory MAGs encoded the potential for

production of trimethylamine through *cutC* or a choline dehydrogenase, including: *Desulfacinum* (M1-7-2-bin.16), *Peptococcia* (M1-7-4-bin.22), and *Desulfitibacterales* (K-7-2-bin.50). Additionally, only two of these MAGs – both inferred SRB - encoded genes for demethylation, *Desulfacinum*, and *Desulfitibacterales*. Both genomes encoded functional potential for demethylation of trimethylamine to dimethylamine, but only the *Desulfacinum* MAG could potentially transform dimethylamine to monomethylamine. Notably, both these genomes were exclusively detected in the STACK-14 well where significantly higher concentrations of choline were detected. No MAG was able to completely demethylate trimethylamine to ammonia. Overall, we infer that the lower salinity in the STACK samples, coupled with the lack of mechanisms for *in situ* quaternary amine biosynthesis, reduces the selective pressure for specific metabolisms such as quaternary amine cycling that are highly beneficial under the brine-level salinities encountered elsewhere.

Appendix A References

1. Daly, R. A. *et al.* Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nat. Microbiol.* **1**, 1–9 (2016).
2. Vengosh, A. *et al.* The Geochemistry of Hydraulic Fracturing Fluids. *Procedia Earth Planet. Sci.* **17**, 21–24 (2017).
3. Shaffer, D. L. *et al.* Desalination and reuse of high-salinity shale gas produced water: Drivers, technologies, and future directions. *Environ. Sci. Technol.* **47**, 9569–9583 (2013).
4. Akob, D. M., Cozzarelli, I. M., Dunlap, D. S., Rowan, E. L. & Lorah, M. M. Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Appl. Geochem.* **60**, 116–125 (2015).
5. Oren, A. Life at High Salt Concentrations. in *The Prokaryotes* (eds. Rosenberg, E., DeLong, E., Lory, S., Stackebrandt, E. & Thompson, F.) 421–440 (Springer Berlin Heidelberg, 2013).
6. Lanyi, J. K. Salt-dependent properties of proteins from extremely halophilic bacteria. *Bacteriol. Rev.* **38**, 272–290 (1974).
7. Oren, A. Bioenergetic Aspects of Halophilism WHY CERTAIN PHYSIOLOGICAL GROUPS OF MICROORGANISMS ARE ABSENT IN. *Society* **63**, 334–348 (1999).
8. Kempf, B. & Bremer, E. Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Arch. Microbiol.* **170**, 319–330 (1998).
9. Sleator, R. D. & Hill, C. Bacterial osmoadaptation: The role of osmolytes in bacterial stress and virulence. *FEMS Microbiol. Rev.* **26**, 49–71 (2002).
10. Whatmore, A. M., Chudek, J. A. & Reed, R. H. The effects of osmotic upshock on the intracellular solute pools of Bacillus subtilis. *J. Gen. Microbiol.* **136**, 2527–2535 (1990).
11. McLaggan, D., Naprstek, J., Buurman, E. T. & Epstein, W. Interdependence of K+ and glutamate accumulation during osmotic adaptation of Escherichia coli. *J. Biol. Chem.* **269**, 1911–1917 (1994).
12. Kuhlmann, A. U., Bursy, J., Gimpel, S., Hoffmann, T. & Bremer, E. Synthesis of the compatible solute ectoine in Virgibacillus pantothenticus is triggered by high salinity and low growth temperature. *Appl. Environ. Microbiol.* **74**, 4560–4563 (2008).

13. Bursy, J., Pierik, A. J., Pica, N. & Bremer, E. Osmotically induced synthesis of the compatible solute hydroxyectoine is mediated by an evolutionarily conserved ectoine hydroxylase. *J. Biol. Chem.* **282**, 31147–31155 (2007).
14. García-Estepa, R. *et al.* The ectD gene, which is involved in the synthesis of the compatible solute hydroxyectoine, is essential for thermoprotection of the halophilic bacterium Chromohalobacter salexigens. *J. Bacteriol.* **188**, 3774–3784 (2006).
15. Bursy, J. *et al.* Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by Streptomyces coelicolor A3(2) in response to salt and heat stresses. *Appl. Environ. Microbiol.* **74**, 7286–7296 (2008).
16. Wood, J. M. *et al.* Osmosensing and osmoregulatory compatible solute accumulation by bacteria. *Comp. Biochem. Physiol. - Mol. Integr. Physiol.* **130**, 437–460 (2001).
17. Borton, M. A. *et al.* Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6585–E6594 (2018).

Appendix B: Chapter 2 Supplemental Data

See supplemental file: AppendixB.zip

Supplementary Data 1. (xlsx) Chemical composition of the frack fluids used in all wells, as disclosed in FracFocus, the national hydraulic fracturing chemical disclosure registry.

Supplementary Data 2. (xlsx) Metabolite and conductivity data. Drill muds do not have metabolite data due to the solid nature of drill mud samples. Conductivity measurements were only taken on produced fluids.

Supplementary Data 3. (xlsx) Metagenomics sequencing and assembly information, as well as samples, including replicates, that also received 16S rRNA gene sequencing.

Supplementary Data 4. (xlsx) Relative abundances and genome statistics of the 24 dominant and persisting STACK metagenome assembled genomes. Relative abundances were calculated with 13Gbp rarified read recruitment to MAGs, following requirements outlined in methods.

Appendix C: Chapter 3 Supplemental Data

See supplemental file: AppendixC.zip

Supplemental Data 1. (xlsx) Sequencing details. Related to STAR methods. (A) Metagenome sequencing and assembly statistics (B) 16S rRNA amplicon sequencing details.

Supplemental Data 2. (xlsx) Details on MAGs. Related to Figure 4 and STAR methods. (A) Information and statistics on MAGs described in this study (B) other viral defense systems detected for all MAGs (C) CRISPR-Cas type classifications from CRISPR-Cas Typer.

Supplemental Data 3. (xlsx) Details on vMAGs. Related to Figure 6 and STAR methods. (A) Additional information and statistics for all 2,176 vMAGs (B) host-viral linkages for MAGs and vMAGs in the new and (C) established wells (D) probable viral lifestyle predictions from BACPHLIP.