DISSERTATION

THE WHOLE IS GREATER THAN THE SUM OF THE PARTS: PIECING TOGETHER MICROBIAL METHYLATED AMINE METABOLISM

Submitted by

Mikayla A. Borton

Department of Soil and Crop Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2020

Doctoral Committee:

Advisor: Kelly C. Wrighton

Michael J. Wilkins
Thomas Borch
Siu Hung Joshua Chan

ABSTRACT

THE WHOLE IS GREATER THAN THE SUM OF THE PARTS:

PIECING TOGETHER MICROBIAL METHYLATED AMINE METABOLISM

Microbial metabolism of methylated amines (MAs), simple nitrogen compounds containing one or more methyl groups, has vast impacts across the globe including mediating greenhouse gas production and human health. In the last decade, new reactions in the microbial MA cycle have been identified and biochemically characterized. While these detailed studies delivered key knowledge of previously elusive MA enzymes, there had not been any studies that collectively interrogated these separate reactions in a holistic manner. The overarching aim of this dissertation was to piece together MA metabolism into a framework that could be applied across ecosystems and ultimately expose implications for this microbial metabolism at an ecosystem level.

To begin to uncover the prevalence of this metabolism, I first needed to summarize what is known about microbial MA metabolism. Thus, Chapter 1 summarizes the key biochemical reactions and enzymes that make up the known microbial MA metabolic network. I also explain in this chapter, how in order to mine these metabolisms from metagenomic datasets, I needed to overcome annotation bottlenecks. It is the combination of newly discovered enzymes, and their poor annotation, that likely explains why these critical processes (e.g. how microbes adapt to deep shales or contribute to cardiovascular disease) remain so enigmatic. It is my hope that my annotation framework will provide a roadmap for mining MA metabolism from genomic datasets. Lastly, the experimental design rational for the data presented in this dissertation is

described in Chapter 1, highlighting how cultivation-based investigations coupled to high-resolution meta-omics at laboratory and field scales was used to tease apart these previously cryptic microbial metabolisms from two environments.

In Chapter 2, I go on to investigate MA metabolism in hydraulically fractured shales. Briefly, hydraulic fracturing of shale is the industrial process behind the surging natural gas output in the United States. This technology inadvertently creates an engineered microbial ecosystem thousands of meters below Earth's surface. To define the MA metabolic network in hydraulically fractured shales, my thesis research first focuses on the methylotrophic methanogens prevalent across geographically distinct shales (Chapter 2). Understanding the metabolism of these methane-producing archaea is necessary to explore the possibility of biostimulation strategies that may be used to increase energy recovery and longevity in hydraulically fractured shale wells, analogous to approaches used on methanogens in coal beds. Next in chapter 3, I scale up MA analyses to the entire microbial community, presenting how pervasive the cycling of these compounds is to persisting shale taxa (Chapter 3). Collectively, Chapters 2-3 illuminate the how microorganisms are utilizing and exchanging MAs for osmoprotection and energy, as well as carbon and nitrogen sources, knowledge that can inform industrial practices for management of microorganisms in hydraulic fracturing ecosystems.

Specifically, in Chapter 2, the biogeography of methanogens across hydraulic fracturing wells is investigated, finding that members of genus *Methanohalophilus* are recovered from every unconventional reservoir sampled to date by metagenomics. Notably, this organism is a methylotrophic methanogen, producing methane from MA compounds such as trimethylamine (TMA). Here, we provide the first genomic sequencing of three isolate genomes, as well as two metagenome assembled genomes within the genus *Methanohalophilus*. Utilizing six other

previously sequenced genomes, we perform comparative analysis of the 11 genomes representing this genus. This genomic investigation revealed distinctions between surface and subsurface derived genomes that are consistent with constraints encountered in each environment. Genotypic differences were also uncovered between isolate genomes recovered from the same well, suggesting niche partitioning among closely related strains. These genomic substrate utilization predictions were then confirmed by physiological investigation. Moreover, fine-scale microdiversity was also observed in CRISPR-Cas systems of *Methanohalophilus*, with genomes from geographically distinct unconventional reservoirs sharing spacers targeting the same viral population. Findings in Chapter 2 not only provides insight into the genotypic differences between strains of the prevalent genus *Methanohalophilus*, but also shows that coupled physiological analyses provided new information on growth parameters that could not have been inferred from genomics alone. Notably, this approach defined different substrate utilization patterns between two closely related and co-occurring strains of this field relevant methanogen, knowledge that is necessary when considering manipulating methanogenic communities in the deep biosphere.

While Chapter 3 also focuses on hydraulic fracturing, it goes beyond a single genus to the microbial community scale, and ultimately aims to define the MA-utilizing members that support shale-relevant *Methanohalophilus*. We used metagenomic, metabolite, and cultivation methods to investigate microbial metabolisms in fluids collected from 5 wells in the Utica and Marcellus shales. Community analyses showed that concomitant with increasing salinities in these wells, persisting microbial communities converged to a similar membership and structure over time, despite differences in operators, shale formations, and input communities/chemistries. Based on the detection of the osmoprotectant glycine betaine (a MA compound) across produced fluids, we

hypothesized that this metabolite may be synthesized *in situ* to support microbial adaptation to increasing salinities.

Based on field data, I designed laboratory reactors to manipulate and monitor persisting shale microbial communities that are currently not feasible in field scenarios. These reactors revealed not only that glycine betaine was synthesized *in situ* by *Methanohalophilus*, but also that other members of the community (*Halanaerobium*), could ferment it to trimethylamine (TMA) a substrate for methanogenesis. Metaproteomic and metabolite findings from the laboratory were then corroborated using regression-based modeling performed on field metagenomic and metabolite data from more than 40 produced fluid samples from five hydraulically fractured shale wells. Collectively, my thesis research showed that *Halanaerobium*, *Geotoga*, and *Methanohalophilus* strain abundances predicted a significant fraction of carbon and nitrogen metabolites at the field scale. Combined laboratory and field results revealed that microorganisms persisting in hydraulically fractured shales must maintain osmotic balance in hypersaline fluids, gain energy in the absence of electron acceptors, and acquire carbon and nitrogen to synthesize cell building blocks. This research provided evidence that co-fermentation of amino acids and their derivatives, like glycine betaine, meets these organismal needs and thus Stickland fermentations function as a keystone metabolism conserved across hydraulically fractured shale communities. Scaling these results from the laboratory to the field identified mechanisms underpinning biogeochemical reactions, yielding knowledge that can be harnessed to potentially increase energy yields and inform management practices in hydraulically fractured shales.

I became curious on the extent of methylamine metabolism across ecosystems. As such, Chapter 4 focuses on the human gut, identifying microbial MA reactions that contribute to cardiovascular disease in humans, centered around metabolite trimethylamine (TMA), the same

key MA leading to methanogenesis in hydraulically fractured shales. Trimethylamine, a nitrogenous metabolite produced by the gut microbiome, is a precursor to trimethylamine-N-oxide, a known promoter of cardiovascular disease in humans. To understand the microbial metabolisms contributing to the microbial formation of this atherogenic metabolite, we built a Gut-Associated Methylated Amine database (GAMAdb) from 238,530 metagenome assembled and isolate genomes, identifying 8,721 genes in 6,341 genomes from 13 phyla that encode methylated amine (MA) metabolism. GAMAdb was coupled to metaproteomics of MA fed gut laboratory reactors that not only show an intricate metabolic network of MA utilizers, but also confirm activity of quaternary amine degradation by gut-derived microorganisms. Moreover, pairing GAMAdb to fecal metagenomic samples from 218 individuals with atherosclerotic cardiovascular disease and 187 healthy controls revealed that MA gene diversity and abundance predicted human cardiovascular disease. Together, generation and application of GAMAdb expanded the diversity of this disease-relevant metabolism, as well as provided a resource for predicting cardiovascular disease from the gut microbiome. Reinforcing the importance of the connection between our microbiota, our metabolites, and our health, this chapter uncovers previously unrecognized players in the gut MA network that are collectively predictive of cardiovascular disease.

The final chapter of this dissertation summarizes the key findings from the hydraulically fractured shales and the human gut. Furthermore, over the course of this dissertation, the relevance of this metabolism in other terrestrial ecosystems has been brought to light using the framework developed in Chapter 1, thus Chapter 5 also aims to summarize the key findings regarding MA metabolism in other terrestrial ecosystems.

In summary, the aims of this dissertation were to summarize the biomarkers and mechanisms underpinning MA cycling in the environment (Chapter 1), uncover MA metabolic

networks in two disparate ecosystems (Chapters 2-4), and summarize the prevalence of this metabolism across several terrestrial ecosystems (Chapter 5). Cumulatively, this dissertation examines MA metabolism in two ecosystems, across multiple scales, to identify the microorganisms and enzymes catalyzing these critical, yet previously cryptic processes undermining microbiome function.

ACKNOWLEDGEMENTS

The culmination of this dissertation becomes a reality with the generous support and guidance of many individuals. I would like to extend my sincerest thanks to all of them.

First and foremost, I would like to thank **Dr. Kelly Wrighton**, as without her expert guidance and vision, this dissertation work would not have been possible. Six years ago, I took her environmental microbiology class as an undergraduate that would change my life forever. The monumental discoveries she taught that semester introduced me to the unseen microbial world and motivated me to ask her for a position in her lab as a technician and eventually a PhD student. I wanted a chance to work in that unseen world and she took a chance on me, providing so many opportunities that made me the scientist I am today. I am beholden for the cutting-edge techniques demonstrated in this dissertation, and even more so for the opportunity to learn them directly from her. Beyond these techniques, she taught me the importance of telling the best story with the data in hand, all the while demanding the highest quality science and forging collaborations that pushed the story to the next level. I am forever grateful for all of our conversations; I cherish our Jimmy Johns writing nights, coffee (and red bull) chats, and white board brainstorming sessions. Kelly, words cannot describe my gratitude for how your kindness, intensity, passion for science, and friendship have shaped me into the scientist and person I am today.

I would also like to thank the members of my committee for generously offering their time and guidance throughout the preparation and review of this dissertation. Specifically, I thank **Dr. Thomas Borch** for acting as guide through the PhD process at CSU, stimulating conversations about hydraulic fracturing, and pushing me to consider other microbiome

techniques in my qualifying exam. I also thank **Dr. Joshua Chan**, who served on my committee- our discussions on flux balance analysis transformed my understanding of what can be done with metagenomic data. Lastly, the continuous support of **Dr. Michael Wilkins** over the last five years has shaped me as a scientist. From our first interaction of showing me how to transform metaproteomics data, to brainstorming on shale papers (and figures), to most recently helping me through the interview process at PNNL (and so many other instances in between)- I am so grateful to him for sharing his time and expertise.

Over the course of the last 5 years, I have had the privilege to dive in the mud with the methanogens of Old Woman Creek, to ponder how extreme life can really get in hydraulically fractured shale, to sift through the numerous gut microbes living in the mouse intestine looking for *Salmonella*'s partners, and to piece together the microbial food web of dietary methylated amines in the human gut. For each of these endeavors, I was a part of an incredible team that provided me with training and background knowledge. From each of you, I have also learned the value of collaboration and the ability to work as a part of a team to answer common research questions.

Wetlands team: I started in the Wrighton lab as a technician working on the wetlands project. Being new to environmental microbiology, I learned so much from everyone on this team, making this project an integral piece of my success in graduate school. Specifically, **Dr. Jordan Angle** not only taught me technical lab skills from doing DNA extractions, to culturing anaerobes, to running the GC, but also answered all of my newbie questions including those that eventually gave me the courage to pursue my own PhD. He also showed me how to lead a team in the field- his composure and professionalism are an inspiration. In addition to Jordan, I thank **Dr. Garrett Smith**, who not only is a fantastic scientist, but is uniquely creative- I thank you for

your brotherly nature and every peculiar food I got to try because you ordered it a lab dinner. More recently, on the wetlands project I have been fortunate to work with **Dr. Adrienne Narrowe**, whose precision and expertise is inspiring. I learn something new from her with every conversation and thoroughly enjoy our sarcastic slack banter.

Salmonella team: I am fortunate to be a member of this collaboration with The Ohio State University. Specifically, I thank **Dr. Brian Ahmer** for allowing me to work in his laboratory on numerous occasions and educating me on the details of *Salmonella* pathogenesis. Within the Ahmer lab, I am also particularly grateful to **Dr. Anice Sabag-Daigle**, who taught me how to work with mice, spent countless hours on microbiome sample collection, and explained many aspects of *Salmonella* pathogenesis, all of which made this project possible. Lastly, I would like to thank **Dr. Vicki Wysocki** whose expertise in metabolomics is unparalleled- I am grateful for her time invested in our shared publications. Lastly, the group meetings held by this team meant so much to my development as a scientist and speaker- thank you all.

Shale team: I am so grateful to have been a small piece of the shale team. This team was an exceptional collaboration of experts spanning several disciplines including **Dr. Kelly Wrighton**, **Dr. Mike Wilkins**, **Dr. Dave Cole**, **Dr. Tom Darrah**, **Dr. Paula Mouser**, and **Dr. Shika Sharma**. The work presented in this dissertation would not be possible without contributions from each of these PIs and their lab groups. I thank each of them for their contributions. In the Wilkins lab, I would specifically like to thank **Dr. Anne Booker**, I appreciate the time she spent showing me how to grow anaerobes under pressure and the fruitful discussions on our shared publications. In the Cole lab, I would like to thank **Dr. Julie Sheets** and **Dr. Sue Welch** for their efforts on sample collection and geochemical measurements. To

Kaela Amundson, there is no other person I can imagine handing shale off to- thank you for being a great colleague and friend.

Methylamines team: Lastly, I would like to thank the human gut methylamine team. **Dr. Joseph Krzycki** and **Duncan Kountz**, thank you for your expertise and guidance on methylamine metabolism, without it much of this dissertation would not be possible. To **Dr. Ruisheng Jiang**, thank you for your contributions to Chapter 4 of this dissertation, specifically the time you spent collecting metabolite data.

I have purposely left **Rebecca Daly** out of the team's paragraphs above because the truth is putting her on one team does not do justice to how she has contributed to this dissertation or any of the projects I have worked on. I am indebted to Reb for the endless hours she spent training me on wet lab experiments, conferring with sequencing facilities, organizing and managing each project, and painstakingly editing everything I have written. Reb is a huge part of the shale team that made much of my work possible including collecting and processing samples, obtaining field knowledge, running laboratory experiments, and training me on viruses. Beyond the science, I cherish our friendship- our take-out lunches in Kelly's office, whiskey dates, and walks to the bridge will not be forgotten. I also plan to take many of Reb's sayings with me in my post-doc including "that figure looks cartoonish" and "we don't sit in science." Thank you, Reb, you have supported me beyond measure.

Outside of academic collaborations, I thank **Dr. David Hoyt** at Pacific Northwest National Laboratory for his NMR expertise and contributions. Specifically, I would like to thank him for his time on many project calls, his explanations of methods, and his numerous searches for MA metabolites of interest - his data and expertise has enabled a deeper understanding of MA metabolism in our datasets for which I am thankful. I also thank **Dr. Mary Lipton** and her

team at Pacific Northwest National Laboratory for her metaproteomics expertise and contributions. Specifically, to **Carrie Nicora**, I appreciate your efforts in metaproteome extraction and spectral analysis, and to **Sam Purvine**, I am thankful for your detailed explanations and many database searches for each dataset.

I would also like to thank all of my other lab mates- having you as friends and colleagues means the world to me. To **Dr. Lindsey Solden**, I would not be here without your initial guidance. Thank you for taking me under your wing and getting me started in this lab- I would not be where I am today without you. To **Dr. Mike Shaffer**, I have enjoyed being your side-kick on DRAM, valued your work on the human methylamines project, and appreciated every time you told me I was sharing my screen on zoom calls. Thank you for being my office mate, while short-lived, I learned so much from you in those six months. To **Bridget McGivern**, thank you for the frequent edits, 14-ners, and wine nights. To **Josue Rodriguez-Ramos**, I appreciate your organization, curiosity, and all of the perfectly timed and hilarious slack memes. To **Kai Leleiwi** and **Katherine Kokkinias**, I have enjoyed being your mentor and watching your progress- you have both been a pleasure to work with.

Most importantly, thank you to my family. To **Tim** and **Karrie Jackson**, my mom and dad, you have been the foundation to my success in so many ways. First, thank you for always encouraging me to follow my heart and never letting me think I couldn't be anything I wanted. Next, thank you for showing me what hard work looks like- it pays off. Lastly, thank you for your continuous kindness and generosity. You have both given me extraordinary examples of how to live, persevere, and be authentic- words cannot express my gratitude for your endless support and guidance. To **Tyler Borton**, I share this success with you. It would not have been possible without your unending support, encouragement, and understanding. To **Margo Khalili**,

you taught me so much in the time you were here, but the most important has to be to fill your life with adventures. This dissertation represents a grand adventure, through many ecosystems, with many people, and out of it comes many stories- I am forever grateful for our friendship.

DEDICATION

This dissertation is dedicated

To Tim and Karrie Jackson who are my compass for everything

&

To Tyler Borton whose unfailing support has

made the last 5 years possible

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1 Methylated amine compounds

Methylated amines (MAs) are nitrogen containing compounds with one or more methyl group(s), with primary, secondary, tertiary, and quaternary classifications being assigned based on the number of carbons bonded directly to the nitrogen atom (Figure 1.1A). These compounds are found across a range of environments and are implicated in catalyzing key biotic functions in plants, animals, and microbes. For example, glycine betaine is a well-characterized osmolyte that accumulates in plant tissue for protection from environmental stresses such as, drought, salinity, UV radiation, and extreme temperatures (1–3). While choline can also function as an effective osmolyte in plants, it is an essential nutrient in humans, functioning as a precursor to phosphatidylcholine, a major component of neuron membranes and shown to have effects on neurocognitive function (4). Likewise, trimethylamine-N-oxide, also has implications on human health, including promotion of atherosclerotic cardiovascular disease in humans (5–7). Moreover, MAs have also been shown to be prevalent in brackish and saline environments, where they originate from marine organisms (8, 9). These compounds also have global implications in terms of climate change, as all levels of methylated amines, including glycine betaine, trimethylamine, dimethylamine, and monomethylamine (Figure 1.1A) are recognized substrates for methanogenesis (10–14).

Given the ubiquity of MAs across Earth's biomes (5, 8, 9, 14–19), it is important to consider the impact of microorganisms on these metabolite pools. Microorganisms, like plants and animals described above, use MA compounds for a variety of functions (Figure 1.1B). Foremost, these compounds contain carbon and nitrogen, which are essential nutrients for all

living organisms and are required for the biosynthesis of key cellular components, such as proteins and nucleic acids (Figure 1.1B). Similar to plants, microorganisms also utilize MAs as osmoprotectants, whereby an organic molecule (e.g. MA like glycine betaine) is accumulated within the cell to maintain osmotic pressure while pumping out the salt (15, 20, 21). Utilization of MAs as osmoprotectants occurs by either uptake from the environment or *de novo* synthesis (15, 20, 21). Lastly, these metabolites represent an energy source for bacterial and archaeal lineages (Figure 1.1B), being utilized in a variety of microbial metabolisms including Stickland fermentation, methanogenesis, acetogenesis, and respiration (11, 14, 22–26).

1.2 Mechanisms of microbial methylated amine metabolism

Microbial transformations of MAs are made up of four key reactions types, including (i) demethylation (pyrrolysine-containing), (ii) demethylation (non pyrrolysine-containing), (iii) MA cleavage, and (iv) MA interconversions (Figure 1.2). Although these reactions are interconnected to create an overall metabolic network (Figure 1.2), the publications incorporating methylated amine metabolisms often only consider a single reaction type in microbiome studies. For example, microbiome studies in the human gut often only consider MA cleavage reactions leading to trimethylamine (e.g. choline TMA lyase, *cutC*), even though demethylation reactions are well documented to also occur in these habitats (27–30). Reasons for lack of integration may be that many of these genes, especially the demethylating genes, have only been recently discovered. Additionally, there are challenges with high throughput annotation of these genes, largely due to the incorporation of pyrrolysine, which is commonly annotated instead as an amber stop codon in metagenome datasets (described below in section 1.3). Thus, the MA network presented in Figure 1.2 represents the first compilation of these metabolisms into a summary metabolic framework, with each reaction type described below.

Demethylation (pyrrolysine and non-pyrrolysine containing). Demethylation of quaternary amines and trimethylamine are carried out by enzymes belonging to the MttB (trimethylamine amine methyltransferase) superfamily (28). This superfamily consists of pyrrolysine containing and non-pyrrolysine enzymes, yet only those with pyrrolysine can have the enzyme function clearly inferred from gene sequence content. These enzymes can demethylate trimethylamine (TMA) to dimethylamine (DMA) (11, 31). Alternatively, the functional assignment of the non-pyrrolysine containing members cannot be inferred from gene sequence content alone (28). A hypothesis first put forth in Ticak, et al. suggests that non-pyrrolysine containing members demethylate quaternary amines (e.g. glycine betaine, carnitine, choline), and many of these have been demonstrated in the past five years (13, 25, 32). Intriguing biochemical evidence has supported this hypothesis, showing that that non-pyrrolysine containing members can demethylate quaternary amines, such as glycine betaine, carnitine, and proline-betaine (25, 28, 32). Considering the chemistry of the pyrrolysine containing demethylating enzymes, it is possible that these non-pyl and pyl containing enzymes function similarly, where the pyrrolysine forms an adduct with trimethylamine, essentially converting it to a quaternary amine before demethylation (28). These demethylating metabolisms have been described in both bacterial and archaeal lineages, with discoveries of pyrrolysine containing MttB occurring in methanogen *Methanosarcina barkeri*, and non-pyrrolysine containing MttB in acetogen *Desulfitobacterium hafiense*, among others.

Other reactions in the demethylating category are carried out by enzymes like dimethylamine methyltransferase (*mtbB*) and monomethylamine methyltransferase (*mtmB*). Both of these enzymes were discovered in *Methanosarcina barkeri* and contain pyrrolysine to remove a methyl-group from dimethylamine and monomethylamine, respectively (10, 11). Notably,

these demethylation reactions of MAs represent a direct link between carbon and nitrogen cycling, ultimately producing carbon dioxide (CO_2) or methane (CH_4) and ammonia (NH_3) (Figure 1.2). This contribution of methylamine metabolism to methane production in soils and other sediments is increasingly being recognized for its more global importance and broad biological prevalence (14). This highlights how microbial MA metabolism influences carbon and nitrogen biogeochemistry.

MA cleavage. These reactions are characterized as one MA being cleaved into a smaller MA (Figure 1.2). A pertinent example of MA cleavage are the set of defined reactions that convert quaternary amines into trimethylamine. Aerobic TMA-producing reactions are carried out by enzymes CntA (carnitine monooxygenase) discovered in *Acinetobacter baumannii* and YeaW (butyrobetaine monooxygenase) in *Escherichia coli,* while anaerobic enzymes include CutC (choline-TMA-lyase) discovered in *Desulfovibrio desulfuricans* and GrdI (glycine betaine reductase) in *Eubacterium acidaminophilum* (22, 27, 33, 34) (Figure 1.2). Notably, these enzymes are key gene targets in human gut, as trimethylamine produced from these reactions is oxidized by host liver enzymes to trimethylamine-N-oxide, a cardiovascular disease promoting metabolite (5–7).

The reactions in this category can be linked to amino acid metabolism. Particularly the glycine betaine reductase (*grdI*), which is a part of a larger family of enzymes including sarcosine reductase that produces monomethylamine from sarcosine and glycine reductase that produces ammonium from glycine (35, 36). These enzymes can carry out reductions that make up half of a Stickland fermentation, which is the oxidation of one amino acid (or derivative) to the reduction of another amino acid (or derivative) (23, 24, 37). This co-fermentation of amino acids generates energy in the form of ATP via substrate level phosphorylation. This metabolism

has been shown to be present in protein-rich environments (15, 38, 39), underlining another aspect of MA metabolism that is critical to microbial life.

Interconversions. Interconversion of MAs represent reactions that are not energy generating metabolisms but convert one MA to another. A prime example of an interconversion is the synthesis of glycine betaine for osmoprotection. Glycine betaine can be synthesized by microorganisms *de novo* from glycine or choline. The three step pathway from glycine is a series of anaerobic methylation reactions that leads to osmolyte glycine betaine (20, 21, 40). These pathways have been demonstrated in methanogenic archaea, including genus *Methanohalophilus* (41). Glycine betaine production from choline requires a two-step, aerobic pathway with a betaine aldehyde intermediate, that has been characterized in genus *Halomonas* (42). Similar to the interconversion of choline and glycine betaine, carnitine can be converted to butyrobetaine with a croto-betaine intermediate (43).

1.3 Methylated amine annotation bottlenecks

Although microbial MA metabolism has been implicated in important ecosystem processes (5, 7, 14, 15, 44), efforts to computationally inventory MA genes is hampered by the lack of characterization in public databases (25, 28, 32), high amino acid sequence similarity within each family (27, 28, 35), superfamily members with unknown functions (27, 28, 45), or genes that are truncated due to pyrrolysine (10, 11, 28, 31).

Considering each of these annotation bottlenecks individually illuminates a workflow to decode the MA functional potential harbored in microbiomes. First, a majority of these genes were discovered and characterized within the last decade (25, 27, 32, 46, 47). Prior to this biochemical annotation, these genes were non-specifically annotated in genome datasets (e.g. glycerol dehydratase, or pyruvate formate lyase), reflecting semblance to their other family

members but not their function (27, 45). In fact, this continues today, as the non-pylrolysine quaternary amine methyltransferases are misannotated as trimethylamine methyltransferases in genomes. Or, in another example, the glycine betaine reductase shares a high sequence similarity with glycine and sarcosine reductases (34–36) and has a bit score of >200, which is well above the default parameters for most automated annotation pipelines (48–50). As such sequence homology alone is not sufficient for annotating these gene types and requires extensive manual curation, likely explaining the absence of these genes in many genome databases.

Simply stated, our current, growing knowledge of microbial biochemistry outpaces our ability to update and disseminate this content into public databases, thus hindering rapid annotation of emerging metabolisms (45, 49, 51). To combat this element of MA gene annotation, I have built custom databases of MA genes to perform targeted searches (Figure 1.3). To enable more accurate annotation of these genes, phylogenies of each family should be built (e.g. glycyl radical enzyme family for *cutC*, glycine reductase family for *grdI*, etc.) and active residues confirmed to assign substrate specificity (27, 35) (Figure 1.3).

While huge strides have been made to define the biochemical functions of MA superfamily members, there are still large branches of each superfamily phylogenetic tree that have unknown functions (28, 45). For example, the MttB superfamily is made up of pyrrolysine containing and non-pyrrolysine methyltransferases (11, 31), of which we can only assign substrates to pyrrolysine containing methyltransferases that demethylate trimethylamine to dimethylamine. The presence of pyrrolysine indicates the position of an amber stop codon, which causes genes to be truncated during automated gene calling. The read-through of this commonly annotated stop codon, and repurposing to pyrrolysine, is another step that needs manual curation of MA genes in genomic datasets (Figure 1.3).

Throughout my thesis, I have identified these gene fragments and manually linked their other halves. Next, I briefly describe this workflow. To differentiate between non-pyrrolysine and pyrrolysine containing homologs in the MttB superfamily, sequences were filtered by length and aligned to known MttB superfamily members. Sequences that were longer than 360 and aligned through the pyrrolysine residue (e.g. not truncated at the pyrrolysine) are non-pyrrolysine containing members. The remaining truncated genes (e.g. break before where pyrrolysine should be) indicate these genes should be reannotated using the amber read-through detection (Figure 1.3). The resulting sequences are pyrrolysine containing MttBs. This same approach should be used for defining functions of monomethylamine and dimethylamine methyltransferases, which also contain pyrrolysine residues (10, 11, 28, 31). Lastly, the MA metabolisms defined in this dissertation also included additional measures to achieve beyond homology based annotation alone (e.g. cultivation of isolates, metaproteomics, and metabolomics).

1.4 An ecosystem perspective on microbial methylated amine metabolism

This body of research uses cultivation-based investigations, coupled to high-resolution multi-omics to interrogate microbial methylated amine metabolism. Specifically, the overarching questions addressed in this dissertation are what is the diversity, prevalence, and interconnectedness of methylated amine metabolic networks and how might these impact overall ecosystem processes. To begin to address these questions, we sampled microbial communities in deep hydraulically fractured shales and the human gut to build representative metabolic networks that were the basis for predictions of ecosystem chemistry and function using multi-omic methodologies. These methodologies were used across different scales of complexity, increasing from isolates and microcosms from one well to field scale investigations across multiple wells in disparate shale formations to a different ecosystem altogether, the human gut. Notably, this

approach highlights the value of scalable and translational studies that was required for deciphering this previously cryptic methylated amine microbial metabolic network. The three primary objectives of my dissertation research were to:

1. Examine metabolic potential and physiology of genus *Methanohalophilus*, a prevalent methylotrophic methanogen in hydraulically fractured shale (Chapter 2) (52).
2. Resolve microbial interactions that underpin persistence in hydraulically fractured shales (Chapter 3) (44).
3. Uncover the diversity and prevalence of microbial methylated amine metabolism in the human gut to predict cardiovascular disease in humans (Chapter 4).

The final chapter summarizes the findings from hydraulically fractured shales and the human gut together with other terrestrial ecosystems, ultimately highlighting how microbial methylamine metabolism may play unrecognized roles in carbon and nitrogen turnover.

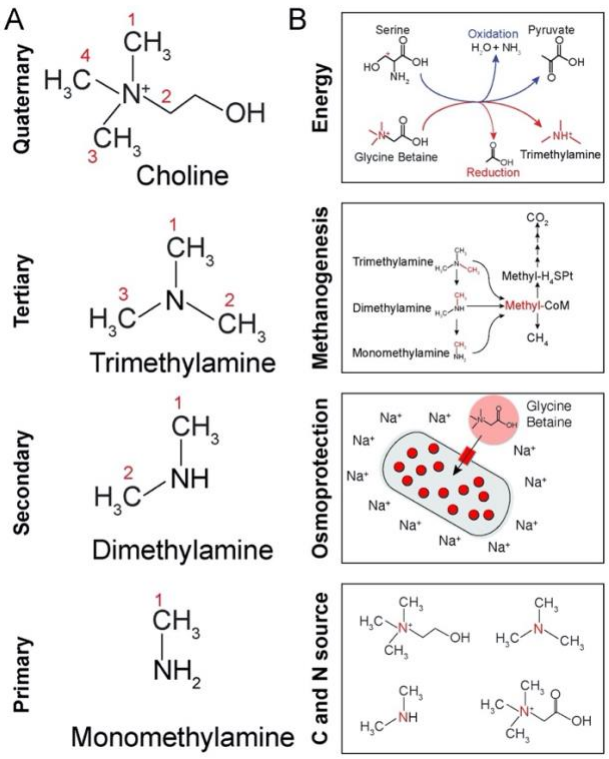

Figure 2.1 Methylated amine structure and importance.

A) Levels of amine (primary, secondary, tertiary, and quaternary) are shown, with methylated amine structures given as examples of each. B) Boxes show key roles that that methylated amines play in microbial ecosystems.

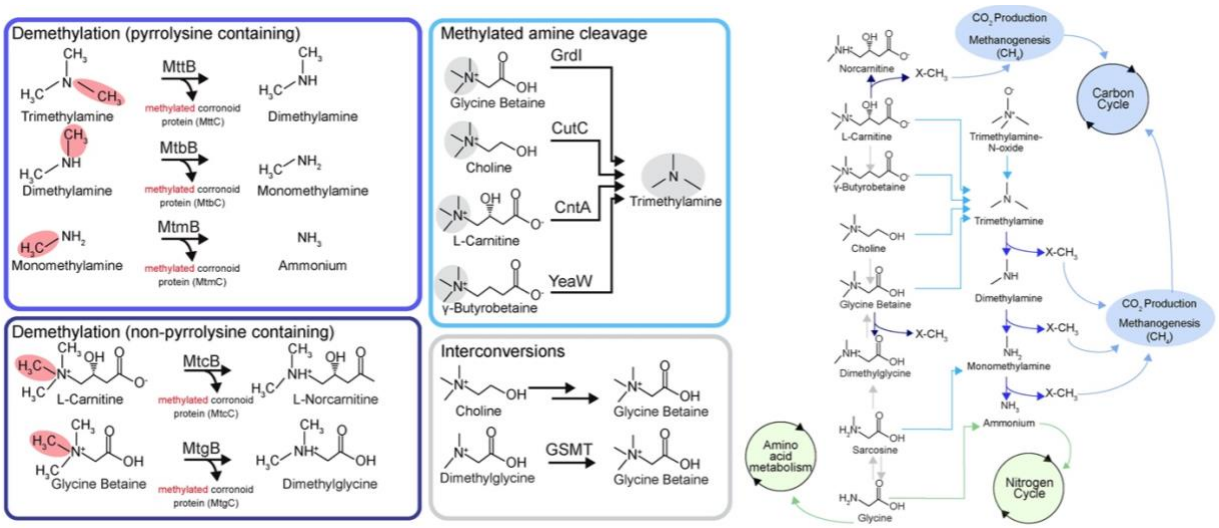

Figure 2.2 Pathways for microbial methylated amine metabolism.

On left, colored boxes denote categories of methylated amine metabolism. Arrows follow the box color scheme in the overall pathway summary (on right). Overall pathway summary shows how methylated amine metabolism integrates together into one network that ultimately feeds into carbon or nitrogen cycles.

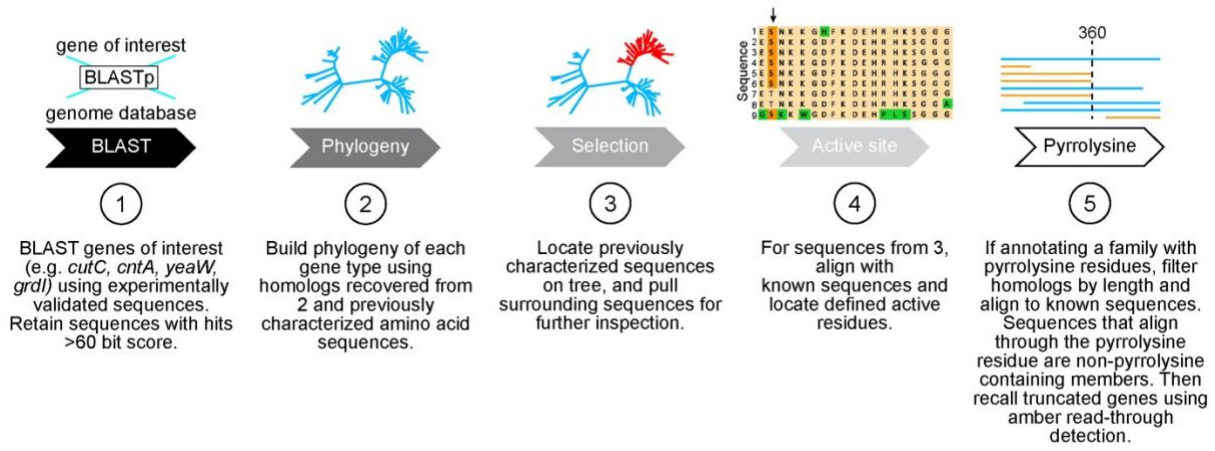

Figure 2.3 Workflow for annotation of methylated amine metabolism.

Annotation of methylated amine metabolism in microbial genomes is shown, with each arrow indicating key steps in the annotation process that overcomes the bottlenecks associated with surveying this metabolism in multi-omic datasets.

Chapter 2: Comparative genomics and physiology of the genus *Methanohalophilus*, a prevalent methanogen in hydraulically fractured shale[1]

2.1 ***Introduction***

Hydraulic fracturing has substantially increased hydrocarbon recovery from unconventional reservoirs, including black shales (siliceous fine-grained mudstones dominated quartz, clays, and carbonates) and micrites (fine-grained organic-rich or "muddy" carbonates with a range of clay content and minor quartz) (53–57). Black shales and micrites, which are commonly classified geologically as mudstones, are thought to be devoid of microbial life following geological "pasteurization" associated with basin loading, burial, and geothermal heating that leads to catagenesis (58, 59). The process of hydraulic fracturing, however, brings life to these unconventional reservoirs by introducing a myriad of biogeochemical, hydrogeological, and structural changes that allow microorganisms to colonize the deep subsurface (59).

Characterized by its name, hydraulic fracturing is the high-pressure injection of water, proppant, and chemical additives into the subsurface that creates fractures in the rock matrix, subsequently releasing economically important hydrocarbons (57, 59, 60). Injected fluids act as a microbial inoculum, transporting surface microorganisms to the deep subsurface. Concomitantly, these injected fluids contain biocides and stabilizers that act as a source of substrates for injected microorganisms (59, 61). These structural and chemical changes create the space and resources

[1] This chapter was reproduced verbatim from "Borton, et al. Comparative genomics and physiology of the genus *Methanohalophilus*, a prevalent methanogen in hydraulically fractured shale. *Environmental Microbiology* (2018)". 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.

necessary for microbial life. Previous studies by our group and others have shown that a subset of members from injected microbial communities persist more than 300 days after hydraulic fracturing and even proliferate in these systems (15, 44, 62).

Microorganisms injected into natural-gas wells invariably have an impact on the deep biosphere environment. The metabolic activity of some persistent microorganisms in unconventional reservoirs can lead to sulfide production, resulting in souring and infrastructure corrosion (63, 64), while the accumulation of microbial biomass may lead to clogging of fractures (65). Other persistent microbial community members include methane producing archaea, which could have a positive impact on energy yields. One prior study predicted that biogenic methane accounted for more than 12% of methane produced in a shale-gas well lifetime (66). Similar to stimulation of methanogens in coal beds (67–69), there is potential to enhance methanogenic metabolism in hydraulically fractured unconventional reservoirs to increase methane recovery and well longevity. In order to make this potential a reality, a deeper understanding of biotic and abiotic processes within the subsurface ecosystems associated with unconventional reservoirs is needed to manipulate and manage methanogenesis.

Here, we survey the biogeography of methanogens across hydraulically fractured unconventional reservoirs, including shales, and perform a comparative analysis of 11 genomes from the most prevalent methanogen, *Methanohalophilus*. This genomic investigation of *Methanohalophilus* strains living in surface and subsurface environments provides insight into the genotypic differences between strains under different environmental constraints. Predicted metabolic differences between strains were supported by laboratory cultivation experiments using *Methanohalophilus* strains isolated from unconventional reservoir produced fluids. Understanding how these methanogens are carving out a life in these dynamic, subsurface

environments is imperative to move forward in field manipulations of their metabolisms for human benefit.

2.2 Results and discussion

2.2.1 Methanohalophilus is a prevalent member of persisting microbial communities in hydraulically fractured unconventional reservoirs

Given the potential economic importance of methane producing archaea to the recovery of natural gas from unconventional reservoirs, such as shale, we surveyed publicly available sequencing data (single gene and metagenomic) from produced fluids for the presence of methanogens (Table 2.1). We combined data from prior published reports spanning five different unconventional energy plays, both from our group (15, 44) and other studies (62, 63, 70–79). We also include new data from the STACK play in Oklahoma (deposited in NCBI Bioproject #PRJNA308326). The unconventional reservoir plays analyzed here represent a variety of different geological histories and geographic locations and thus have variable chemistries and conditions for microbial life.

Due to the bias of universal primers (80, 81), especially for archaea (82), we preferentially utilized metagenomic data when available (Table 2.1). We report the metagenomic data in two ways, i) read mapping to known methanogens to assess relative abundance and ii) metagenome assembled genomes (MAGs) for assessment of metabolic potential (see methods). Here we release five new *Methanohalophilus* genomes, comprised of three isolates and two MAGs (Table 2.2). For wells analyzed without metagenomics data, especially those from other groups and formations, 16S rRNA gene data was the only available information (Table 2.1). For these, we assessed methanogen presence and diversity, but did not include relative abundance. Together, our analyses include data from 44 natural-gas wells across six unconventional

formations, enabling a biogeographical survey of potentially important methane producing archaea.

To date, every microbial community analysis in hydraulically fractured unconventional reservoirs that included archaea detected the presence of methanogens (Table 2.1, Figure 2.1A). While *Methanohalophilus* (Figure 2.1B) have been found in the six unconventional reservoirs plays sampled to date, other methanogens including *Methanolobus* and *Methanoplanus* have been occassionally detected (15, 76). Consistent with other methanogenic ecosystems(83–85), the relative abundance of methanogens in unconventional reservoirs, including shales, is generally low, typically less than 7% (Figure 2.1C). The cosmopolitan nature of *Methanohalophilus* in source rock formations from the Marcellus, Utica, Barnett, Antrim, and STACK unconventional reservoir plays, hints at a keystone role in these ecosystems.

To better define a "universal" niche occupied by *Methanohalophilus*, we mapped metagenomic reads to *Methanohalophilus* isolate genomes and MAGs to determine the relative abundance of these genotypes in fluids produced from unconventional reservoirs over time. Time series metagenomic data shows that *Methanohalophilus* was detected in samples up to 488 days post-hydraulic fracturing, the latest time point sampled (Appendix A). Previous studies have used 16S rRNA gene data to show that microbial communities in unconventional reservoirs often converge to a similar microbial community at late time points regardless of input chemistry, operator, geographic location, or type of source rock (i.e., shale, organic-rich micrite) (44, 59). *Methanohalophilus* is a part of this persisting community, along with *Halanaerobium*, *Geotoga*, and *Marinobacter* (Figure 2.2).

One hypothesis for this convergence of similar genotypes across natural-gas well samples from different unconventional reservoirs is strong environmental filtering. A likely critical

environmental forcing is salinity, which changes throughout the lifecycle of a natural-gas well. In many cases salinity can range from freshwater (<100 mg/L) up to hypersaline conditions (>110 g/L) over this time period (15, 44, 75). In general, salinity increases in a power function with time, with the shift to saline conditions occurring on a time scale of weeks to months (Figure 2.3). Metagenomic data paired to salinity data shows a shift to saline conditions over approximately 40 days after hydraulic fracturing (Figure 2.3), a change that is reflected in microbial membership (Figure 2.2). Notably, the change in microbial community membership is characterized by an enrichment of halophilic or halotolerant taxa 40 days after hydraulic fracturing.

Here we use metagenomic data to define a salinity window for *Methanohalophilus* across the transition from flowback to produced fluids. Open circles show that *Methanohalophilus* was not detected below 25 g L^{-1} chloride, while closed circles show the upper limit for *Methanohalophilus* in produced fluids is 112 g L^{-1} chloride (Figure 2.1C). The upper limit is also the maximum concentration for this dataset and thus we cannot conclude the maximum chloride concentration tolerated by this genus in hydraulically fractured unconventional reservoirs. This reported range is consistent with a published report indicating the upper limit for methylotrophic methanogenesis (not just *Methanohalophilus*) is 250 g L^{-1} chloride (20). Based on analysis of the literature for cultivated *Methanohalophilus spp.* (41, 86–88), we conclude the salinity range for this genus spans from 17- 155 g L^{-1} chloride, with an optimum of around 80 g L^{-1} chloride (Table 2.3). Here our produced fluid data suggests that there is no clear optimum salinity concentration (Figure 2.1C), and in fact there are several samples containing around 80 g L^{-1} chloride where *Methanohalophilus* was not detected, and low salinity samples (STACK formation) where *Methanohalophilus* was detected.

We suggest several explanations for the variable *Methanohalophilus* distribution. The first is methodological, in that the absence of data at certain time points can be impacted by sampling, i.e., the sequencing depth combined with the lower relative abundance of methanogens relative to other members in the community may result in non-detection of methanogens. Second, there are other chemical and biological factors that can mediate the presence of *Methanohalophilus* in hydraulically fractured unconventional reservoir systems. For example, we have previously shown that methanogenesis by *Methanohalophilus* can be fueled by *Halanaerobium* fermentation of glycine betaine to yield methanogenic substrates such as trimethylamine (15). Consistent with this interdependency, *Halanaerobium* is present in every sample in which *Methanohalophilus* is detected. However, here we failed to find any clear relationship between *Methanohalophilus* presence or relative abundance and the presence or concentration of methylotrophic substrates (Figure 2.4). Lastly, we previously demonstrated that viral genomes could be linked to *Methanohalophilus* (15), and that viral predation and *Methanohalophilus* CRISPR immunity were actively expressed (44); thus predation may also contribute to variable *Methanohalophilus* abundance patterns. We conclude that *Methanohalophilus* requires a specific set of yet-defined conditions (and maybe even microbial members) to persist in deep hydraulically fractured unconventional reservoirs. Developing more detailed knowledge of the constraints on this organism will be required to successfully manipulate this system for enhanced biogenic methane production.

2.2.2 New Methanohalophilus MAG and isolate genomes double prior genomic sampling of this genus

For this study, we supplemented publicly available genomes with in-house MAGs (n=2) and isolate genomes (n=3) all belonging to the genus *Methanohalophilus* (Table 2.2). In total, we

compared 11 genomes; six isolates and five MAGs, three of which are in one contig (*Methanohalophilus mahii* SLP DSM 5219, *Methanohalophilus portucalensis* FDF-1, and *Methanohalophilus euhalobius* WG1-MB) (Table 2.2, Appendix A). Three of the genomes are from surface environments (86, 87, 89), including lake sediments (*Methanohalophilus mahii* SLP), salt pans (*Methanohalophilus portucalensis* FDF-1), and cyanobacterial mats (*Methanohalophilus halophilus* Z-7982). The remaining eight genomes are from subsurface environments, with all but one, which is from an oil well in Russia (90), being from hydraulically fractured unconventional reservoirs in the US (15, 44, 71) (Table 2.2).

The 11 genomes range in size from 1.6-2.09 Mbp, with no size differences noted due to differences in isolate versus metagenome-derived genomes (Table 2.2, Appendix A). The genomes encode 1,889-2,233 predicted proteins and have between 41.9-42.6% G+C content. All of the genomes are estimated to be >80% complete with <4.8% contamination (Appendix A). Genomes first described here include three isolates and two MAGs, which nearly doubles the number of sequenced representatives of this genus. All of the genomes analyzed share more than 91% average nucleotide identity (Figure 2.5). Pangenome analysis revealed 3,112 gene clusters among the 11 genomes, with 901 genes constituting the core (present in every *Methanohalophilus* genome). The accessory genome included 1,445 genes present in at least two genomes and 766 unique genes that were present in only one genome (Figure 2.6, Appendix A). The core genes account for 42.5-49.4% of the total genes across the 11 genomes. A phylogenetic analysis constructed with the core genes, as well as standard Archaeal marker genes (91), showed that three different *Methanohalophilus* species (*mahii*, *halophilus*, and *portucalensis* spp.) were sampled from surface habitats (n=3 genomes), while *Methanohalophilus euhalobius* was only sampled in the deep subsurface (n=8 genomes) (Figure 2.5, 2.7).

The percentage of genes in the *Methanohalophilus* core genome is much lower than *Thermotoga* (>90% core genes, n=11), a bacterium that spans surface (n=4) and deeper (n=7) ecosystems (92). In fact, the relative abundance of genes in the core genome for *Methanohalophilus* is closer to *Prochlorococcus*, a highly abundant marine photosynthetic bacterial genus, with a core genome of about 50% (n=49) (93). Within the archaea, a genome comparison of a single genus and species, *Sulfolobus islandicus* (n=7), recovered from surface, geothermal areas revealed 74% of the genes made up the core genome. Here a similar analysis of *Methanohalophilus euhalobius* genomes recovered exclusively from the deep hydrocarbon systems (n=8), had only 46% of the genes in the core genome.

Reasons for the low percentage of genes in the *Methanohalophilus* core genome may be due to the engineered nature of hydraulically fractured rocks in unconventional reservoirs, which are atypical from other subsurface or deep ecosystems. Microorganisms from undisturbed deep biosphere environments often have slow growth rates that may hinder nucleotide substitution accumulation, ultimately keeping diversity low (94). However, previous studies have shown that in hydraulic fractured rocks in unconventional reservoirs, biomass can increase by over several orders of magnitude over a period of several months (15). Moreover, increased genetic diversity within *Methanohalophilus euhalobius* strains relative to other subsurface archaea, like *S. islandicus* may be attributed to environmental instability, which is a key driver of genetic diversity (95). The ecosystem created following hydraulic fracturing is characterized by extreme changes in salinity, redox conditions, temperature, pressure, nutrients, and perhaps even viral predation throughout the lifetime of the natural-gas well (15). It is important to consider that these environmental drivers and the static conditions prior to hydraulic stimulation vary considerably for each unconventional reservoir (44) (Figure 2.3). For example, among

unconventional reservoirs, there are important differences in dominant (e.g. quartz, clay, carbonates) and trace (e.g. sulfides) mineral assemblages, burial depth (0.5 to >3km), reservoir temperature (40 to ~150°C), total organic carbon content (0.5 to ~20%), kerogen type, basin burial history, and geologic history (e.g., occurrence of volcanic intrusions, faulting, fracture intensity, style) (58, 59, 96–103). Together these differences may facilitate the increased genomic diversity we observed within the genus *Methanohalophilus*.

The core genome of *Methanohalophilus* includes the expected housekeeping genes, but also the genes necessary for methylotrophic methanogenesis and production of the osmolyte glycine betaine (Appendix A). For the methylotrophic substrates, only the utilization of methanol and monomethylamine are conserved, while trimethylamine and dimethylamine utilization are flexible genome attributes (Figure 2.7). Specifically, all 11 genomes have the capacity to demethylate methanol and monomethylamine via substrate specific pyrrolysine-containing methyltransferases (*mtaB* and *mtmB*, respectively) that transfer the methyl group to respective activated cognate corrinoid proteins, which are also conserved (*mtaC* and *mtmC*, respectively). Methyl groups are ultimately transferred to coenzyme M via a methylcorrinoid:CoM methyltransferase (*mtbA*), which is encoded in all 11 genomes. Genes for pyrrolysine biosynthetic enzymes (*pylBCD*), methyl coenzyme M reductase (*mcrA*), and a pyrrolysine-tRNA synthetase (*pylRS*) were also recovered from all 11 genomes, as these are required to utilize methylamines. Genes for the utilization of other methanogenic substrates such as acetate or quaternary amines (e.g. glycine betaine, (28)) were not recovered in any of these 11 genomes. This is consistent with prior reports in which *Methanohalophilus* is an obligate methylotrophic methanogen, requiring the presence of methanol or methylamines to generate energy (86, 104).

Given the elevated and increasing salinities in produced fluids through time, we surveyed the *Methanohalophilus* pangenome for the presence of genes involved in osmoadaptation. Notably, genes for the synthesis of glycine betaine from glycine, as well as the transport of glycine betaine from the environment into the cell, is present in all 11 genomes. The synthesis or transport of other osmolytes was not in the core genome for *Methanohalophilus*. This shows the importance of glycine betaine to *Methanohalophilus* physiology. Given that glycine betaine is the only osmolyte to be detected in all produced fluids from late time points, we had previously suggested that this compound was a keystone metabolite in the ecosystem created by hydraulic fracturing (44). Moreover, we demonstrated using metaproteomics data from laboratory reactors that this compound can forge metabolic interactions between persisting taxa. For instance, glycine betaine is synthesized from glycine by *Methanohalophilus*, used as an osmoprotectant by *Methanohalophilus* and *Geotoga*, and used as an energy source by *Halanaerobium* and *Frackibacter* (15, 44, 105). Also, given that glycine betaine is an amino acid derivative ($C_5H_{11}NO_2$), it can serve as a source of organic nitrogen in this system. Similarly, *Methanohalophilus*-fueled glycine betaine metabolisms are also reported in a surface hypersaline lake (17). Thus, the highly conserved capacity to produce of glycine betaine by *Methanohalophilus* may represent a public good, contributing to ecosystem stability in saline, methanogenic ecosystems.

2.2.3 Unique attributes of subsurface Methanohalophilus genomes

It is widely presumed that the flexible part of an organism's genome confers fitness advantages to specific strains within different environmental conditions (93). As such, we examined the flexible genome between surface and subsurface *Methanohalophilus* genomes. There are 83 gene clusters unique to surface genomes and 40 gene clusters unique to subsurface

genomes. Of these, 51% and 45% are hypothetical proteins or genes with unknown function, respectively, alluding to the currently cryptic biochemistry that resides in lineages with well-characterized members (106, 107).

Genes present only in subsurface genomes may have implications for adaptation and persistence in the deep biosphere. For instance, one gene cluster that was only found in subsurface derived genomes was made up of 37 annotated transposases. Although there are other transposase gene clusters in surface genomes, the subsurface genomes contain significantly more copies of transposases relative to surface genomes ($p<0.05$). The relative abundance of transposase genes to total genes in a given genome was 0.3-0.6% for the *Methanohalophilus* surface genome, and 1.0-2.3% for subsurface genomes. Interestingly, this percentage cut-off of 1% is consistent with prior reports showing that surface genomes derived from marine systems typically have 0.6% transposase relative abundance, while "extreme" environments like acid mine drainage have 1% transposase relative abundance (94). The abundance of transposases in genomes from subsurface or extreme environments suggests an active role for genome plasticity and adaptation in environments with strong selective pressures (108).

Other genes exclusive to subsurface genomes include a specific CRISPR-associated protein belonging to the Csx1 family. This particular protein has been previously implicated as an endoribonuclease that acts selectively on single-stranded RNA and cleaves specifically after adenosines (109). We note from a prior proteomic examination of *Methanohalophilus* 2-GBenrich genome that this CRISPR gene was expressed in a laboratory maintained consortium experiencing viral predation (44). Thus, the conservation of this gene may suggest an adaptation of subsurface *Methanohalophilus* against viruses that surface *Methanohalophilus* may not have encountered, a hypothesis we examine in more detail below. Genes for an alcohol dehydrogenase

(ADH) were also found exclusively in the subsurface genomes. While this may confer adaptation to chemical conditions in oil and natural-gas wells, further investigation into the specificity and annotation of this gene is necessary. Notably, genes for glycan production were also found exclusively in subsurface *Methanohalophilus* genomes, potentially associated with a role in biofilm formation in the deep biosphere (110, 111).

Genomes from the surface contain genes that may confer adaptation to unfavorable redox or light conditions that are more common in these habitats. For example, catalase genes were only recovered from surface *Methanohalophilus* genomes. Recent reports from soils have suggested the expression of this gene is associated with oxygen detoxification and can enable methane production in oxic habitats (112). Other genes exclusively encoded in surface genomes included nitric oxide reductases. This is consistent with a prior publication showing that other methanogens may use this gene as a detoxification mechanism (113). Lastly, surface *Methanohalophilus* genomes exclusively contain genes for photo-lyase enzymes, which repair DNA damaged by ultraviolet light. In summary, we identified disparities in genomic content that reflect the differences in environmental selective pressures encountered by these microorganisms in surface (salt pan, cyanobacterial mat, lake sediments) and subsurface (unconventional petroleum reservoir) habitats.

Genes for the utilization of trimethylamine (TMA), a common methylotrophic substrate, was part of the flexible genome. We identified differences in the use of TMA by two closely related strains isolated from the same natural-gas well. In fact, these two organisms were so phylogenetically similar that they were indistinguishable by 16S rRNA gene sequence (100% identify of 1,456 basepairs) and shared nearly 99% average nucleotide identity at the genome level. Specifically, *Methanohalophilus* WG1-DM does not encode the genes for trimethylamine

utilization (corrinoid protein and trimethylamine methyltransferase), while *Methanohalophilus* WG1-MB does contain these genes (Figure 2.7-2.8). The presence of trimethylamine methyltransferase was detected in all other *Methanohalophilus* genomes (n=9 of 11, except the DM isolate and a closely related MAG from the same sample). Other genome differences between these two closely related strains cultivated from the same natural-gas well include an oligopeptide permease ABC transporter operon (opp) and antitoxin system (pemK/mazF). Knowledge of the differential substrate utilization patterns between strains represent necessary information when considering the rational design of biogenic methane-producing communities. Moreover, these inferences cannot be made based on genomes alone, as physiological parameters like growth efficiency and rate also need to be considered when enhancing methanogenesis at the field scale.

2.2.4 Genome strain differentiation validation through physiological investigation

Due to methodological constraints involved in assembly and binning, the absence of genes in isolate and metagenomic derived genomes is not alone sufficient for inferring physiology. To validate the absence of TMA utilizing genes in strain WG1-DM, we performed physiological characterization of these two closely related strains (*Methanohalophilus* WG1-DM and *Methanohalophilus* WG1-MB). Between strains we compared the growth rate, growth yield, and the total methane produced on four separate methylotrophic substrates (Figure 2.8). All of the methylotrophic substrates evaluated here (methanol, trimethylamine, dimethylamine, monomethylamine) were present through time in the produced fluids, demonstrating the relevance of these compounds to the ecology of *Methanohalophilus* (Figure 2.8).

Our physiological data verified the substrate utilization patterns inferred by genomics. For instance, WG1-MB increased to >1 optical density (O.D. measured at 600nm) and produced

31.5 μmol of methane produced over a period of nine days, while no growth or methane production was observed by strain WG1-DM on the same concentration of TMA given the same amount of time. Additionally, physiological analyses provided new information on growth parameters that could not be concluded from a genome. For example, the exponential growth rate of WG1-DM and WG1-MB did not differ when grown on dimethylamine or methanol but was statistically different on monomethylamine. For this latter substrate, WG1-MB had a faster exponential growth rate than WG1-DM, but no significant differences in total cell or methane yield were observed. We note that these enzymes and their corresponding coronoid proteins are identical at an amino acid level. Also, in these experiments, we attempted to account for differences in starting biomass, resting metabolic state, and substrate concentrations, as we used washed cell suspensions with the similar cell density for inoculation. Thus, it is possible that these growth rate differences could be attributed to variations in substrate transport or uptake (e.g. transporter, cell envelope proteins), growth forms (biofilm, planktonic), or currently unknown enzyme or kinetic or pathway efficiencies.

2.2.5 Viral predation in Methanohalophilus leads to strain differentiation

Previous work from our laboratory has shown that viruses are important controllers in unconventional reservoir ecosystems (15, 44). In line with these findings, nine of the eleven *Methanohalophilus* genomes in this study contain CRISPR-Cas systems, which is an acquired immune system used by bacteria and archaea to ward against viruses and other invading foreign DNA. A CRISPR array is a hyper variable region within a bacterial or archaeal genome, composed of direct repeats and spacers, with each spacer recording a successful defense against viral or foreign DNA invasion (114). All *Methanohalophilus euhalobius* genomes, which were sampled exclusively from natural-gas and oil wells, have CRISPR arrays (Appendix A), with a

majority being denoted as type I CRISPR systems (especially type I-A) (114) (Appendix A). Comparison of the CRISPR-Cas arrays revealed 11 direct repeat sequences, which were either 30 or 37 bp in length. Collectively, the *Methanohalophilus* genomes contained a total of 993 spacers in 28 CRISPR-Cas arrays. All but one of the arrays were not at the end of contigs, suggesting these numbers are not methodologically inflated due to assembly breaks. Remarkably, 52% of the collective spacers in *Methanohalophilus* genomes are shared by at least one other *Methanohalophilus* genome (Figure 2.9A, Appendix A), despite differences in reservoir geographic location (ranging from the U.S.A. to Russia), reservoir formation (e.g. Marcellus, Haynesville, and Utica), and reservoir age (spanning from the Ordovician (Utica: ~465-450 Ma) through the Jurassic (Haynesville: ~151- 125 Ma)). Taken together, our findings suggest that *Methanohalophilus* genomes are under similar predatory stress regardless of their geographic location, rock type (ranging from shales (siliceous) to micrites), or depositional age (spanning from ~465-125Ma in this study).

The two highly similar *Methanohalophilus* isolates, WG1-MB and WG1-DM, that originated from the same natural-gas well and were physiologically characterized (see above), share 26 identical spacers in three different CRISPR-Cas arrays (Figure 2.9A). While some of these arrays are identical (WG1-DM CRISPR-4 and WG1-MB CRISPR-4), other arrays contained a series of identical spacers followed a series of divergent spacers. For example, WG1-MB CRISPR-2 and WG1-DM CRISPR-1 share nine consecutive spacer sequences but each have three unique spacers at the end of the array. Similarly, WG1-MB CRISPR-5 and WG1-DM CRISPR-3 share identical first four spacers and the next 13 spacers are divergent (Figure 2.9A). Together these data suggest that in recent evolutionary history, the *Methanohalophilus* isolates WG1-MB and WG1-DM may have constituted a single population that diverged, with each

strain subsequently encountering different viruses. Both of the closely related strains are present during the Utica 2 natural-gas well lifetime, with no differences in relative abundance, thus we could find no fitness advantage by the type of spacers maintained in these genomes (Appendix A). Here we show differences in substrate utilization, growth rate, and perhaps some yet appreciated component of viral immunity may contribute to subtle micro-diversity maintained in the deep subsurface.

Beyond isolates, a comparison of MAG CRISPR arrays allow for microdiversity and viral history characterization across geographic distances. For instance, a comparison of MAGs from two separate Marcellus natural-gas wells (genomes 1-M1, DAL1), operated by different companies and data collected by separate groups, revealed an identical CRISPR arrays spanning 34 spacers. These two genomes also have an array that shares the first 14 spacers, but the DAL1 genome has additional 55 spacers unique to that genome (Appendix A). This suggests while viral populations have a broad host range, others may be natural-gas well or strain specific.

This pattern, of identical and divergent arrays, also holds true when comparing arrays within an isolate from an oilfield in Bonduzkhoe, Russia (*Methanohalophilus euhalobius* DSMZ 10369) and a MAG from a natural-gas well in the Marcellus Formation, sampled in West Virginia, U.S.A (*Methanohalophilus* sp. 4-M4). Eleven spacers are 100% identical (10 consecutive spacers) between these two genomes. Collectively, these data suggest that despite being sampled from different natural-gas wells in the same unconventional formation (i.e., a shale in this instance), different unconventional reservoir formations within the U.S.A., and between the U.S.A. and Russia, *Methanohalophilus* genomes sampled to date have likely encountered the same viral genomes or populations.

The comparison of methanogen spacer sequences to viral genomes assembled and binned from the same samples (15), allowed us to link viral sequences to *Methanohalophilus* genomes. We found that 61 spacers, or ~8% of the total spacers, in the seven *Methanohalophilus euhalobius* derived from hydraulically fractured unconventional reservoirs, matched a single viral genomic population (Figure 2.9B). We note the other non-shale *Methanohalophilus euhalobius* genome (DSMZ 10369), which has a CRISPR system with some shared spacers to these unconventional reservoir genomes lacked spacers for this virus. A genome representing this conserved unconventional reservoir virus was reconstructed, resulting in a circular 54,653 bp genome (15). We mapped *Methanohalophilus* spacers to this viral genome to identify viral genome regions commonly incorporated as spacers in *Methanohalophilus euhalobius*. Studies have suggested these new spacers are organized around protospacer adjacent motif (PAM) (115), thus future identification of *Methanohalophilus* PAMs may continue to improve insight into *Methanohalophilus* viral immunity.

In summary, our comparative genomic and viral analyses lead to several conclusions about *Methanohalophilus* viral-interactions. First, many of our *Methanohalophilus* spacers could not be linked to viral genomes, suggesting we are under-sampling the archaeal viral diversity. Second, the fact that a single viral population had 63 links to *Methanohalophilus* genomes from unconventional reservoirs suggests that this viral genome is broadly distributed within the subsurface, at least in the Appalachian Basin. Moreover, this viral population can infect multiple *Methanohalophilus* strains in and between wells. Additionally, multiple spacer hits (up to 23) in a single methanogen genome to this viral population suggests that the *Methanohalophilus* and their viruses are in an arms-race between host immunity adaptation through CRISPR-Cas spacer incorporation and viral mutation (116, 117). Optimizing methane production in these

economically important ecosystems is likely going to require knowledge not only of methanogen physiology, but also predator-prey interactions.

2.3 Conclusion

We performed the first comparative genomic analyses of a methanogen prevalent in saline ecosystems. We contribute five isolate and metagenome-assembled genomes for the genus *Methanohalophilus*. We highlight genes that may enable adaptation in both surface and deep subsurface habitats. We couple genome predictions to laboratory physiological characterizations to define niche partitioning between two closely related, and co-occurring strains. We provide data that supports our speculation that in hydraulically stimulated, unconventional reservoirs the genome plasticity observed in our *Methanohalophilus* genomes could be maintained both by dynamic changes in environmental conditions, as well as viral predation and transposable elements. These results have implications for manipulating methanogenic communities in the deep biosphere, a rational, ecosystem-based design which could ultimately minimize souring, and lead to enhanced natural-gas production with extended natural-gas well longevity.

2.4 Materials and Methods

2.4.1 Isolation of two Methanohalophilus strains from produced fluid of a Utica 2 natural-gas well

Methanohalophilus WG1-DM and *Methanohalophilus* WG1-MB were isolated from produced fluid samples collected from the same gas–fluid separator 94 and 96 days post hydraulic fracturing, respectively. Each isolation was done using modified DSMZ 479 media dispensed in Balch tubes sealed with butyl rubber stoppers and aluminum crimps under an atmosphere of N2/CO2 (80:20, vol/vol). The modified DSMZ medium (per liter) included 87 g sodium chloride, 1.5 g potassium chloride, 6.0 g magnesium chloride, 0.4 g calcium chloride, 1.0

g ammonium chloride, 2.0 g yeast extract, 2.0 g trypticase peptone, 0.2 g coenzyme M, 0.2 g sodium sulfide, 4.0 g sodium bicarbonate and brought to a pH of 7.2 using 1 mM NaOH. Both strains of *Methanohalophilus* were isolated via serial dilutions on trimethylamine (WG1-MB) and dimethylamine (WG1-DM).

2.4.2 Methanohalophilus imaging

Methanohalophilus WG1-MB cells were imaged (Figure 2.1B) at the Molecular and Cellular Imaging Center, Ohio State University (https://mcic.osu.edu/home). An equal volume of 2x fixative (6% glutaraldehyde, 2% paraformaldehyde in 0.1 M potassium phosphate buffer pH 7.2) was added directly to 1 ml of cell culture. Cells were precipitated and resuspended in 50 uL of 0.1 M potassium phosphate buffer pH 7.2 and applied to a silicon waffle (Electron Microscopy Sciences, catalog # 71893-08). The sample was dehydrated in a graded ethanol series, transitioned into 100% hexamethyldisilazane, and air dried. Images were then obtained with the Hitachi S4700 scanning electron microscope.

2.4.3 Sample collection, geochemistry and metabolite analyses of produced fluids

Sample collection, geochemistry, and metabolite analyses (via NMR) from Utica/Point Pleasant and Marcellus natural-gas wells were reported previously in (44). These samples were from natural-gas wells in Ohio (n = 2), West Virginia (n = 2), and Pennsylvania (n = 1). New samples included here were from Oklahoma wells (n=3) in the STACK play and do not include metabolite analysis by NMR. Reservoirs sampled here range in age of formation: Utica: ~465-450Ma; Marcellus: ~390-365Ma; Antrim: ~375-360Ma; STACK: ~375-350Ma; Barnett: ~350-323Ma; Haynesville: ~151-125Ma.

Chloride concentrations for the new samples provided here from the STACK play were analyzed as previously described in (15). Briefly, chloride concentrations from produced fluids

were obtained using a Thermo Scientific Dionex ICS-2100 ion chromatograph and are included Appendix A.

2.4.4 Sequencing, genome assembly, annotation and binning

All isolate genomes were downloaded from the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes (IMG/M) database. For isolates reported here (n=3, Table 2), DNA was sequenced at the Department of Energy Joint Genome Institute (JGI), Walnut Creek, CA, USA. Illumina shotgun libraries were constructed and sequenced using the HiSeq 2500-1 TB platform. The Illumina sequence data were assembled using the CLC Genomics Workbench (version 8.0.1) and AllPaths-LB (version r46652).

Previously reported MAG scaffolds were downloaded from NCBI using the accession numbers specified in Table 2.2(15, 44). MAGs reported here, including *Methanohalophilus* 3U2 and *Methanohalophilus* 4M4, were binned manually using a combination of GC content, taxonomy and coverage from the Utica 2 natural-gas well Day 94 metagenome (reported in (44), JGI accession number 3300006807) and Marcellus 5 Day 313 metagenome (reported in (44), JGI accession number 3300013017), respectively. Sequencing methods for these metagenomes are described in detail in (44)(the publication in which they were first released). As described previously (15, 118), genome completion was estimated based on the presence of core gene sets (Bacteria, 31 genes and Archaea, 104 genes), using Amphora2 (119). Contamination (gene copies >1 per bin) indicating potential misbins, along with GC and phylogeny, were used to manually remove potential contamination from the bins.

All genomes were annotated as previously described in(15). Briefly, open reading frames were predicted with MetaProdigal (120), and sequences were compared with USEARCH (121)

to KEGG, UniRef90, and InterProScan (122) with single and reverse best hit (RBH) matches of >60 bases reported.

2.4.5 Methanohalophilus relative abundance

Reads from previously published produced fluid metagenomes (44) and 5 additional samples reported here were competitively mapped to a database of 11 *Methanohalophilus* genomes and strain CRISPR arrays (Appendix A) using Bowtie2 (123). Relative abundance was obtained by quantifying the percent of reads that mapped with zero mismatches (the number of reads that mapped divided by the total reads in the metagenome). Presence of *Methanohalophilus* was confirmed by manually looking for *Methanohalophilus* scaffolds (greater than 2 kb) in each assembled metagenome. We note that for one prior publication the reads were not publicly available, only genomic scaffolds, and thus relative abundance of *Methanohalophilus* could not be determined for this sample (71).

2.4.6 Comparative genomics of 11 Methanohalophilus genomes

The 11 *Methanohalophilus* genomes were analyzed using the Anvi'o (version 5) pangenomic workflow (124, 125). First an Anvi'o contigs database was generated using gene calls and annotations from our in-house annotation pipeline, described above and published previously (15, 118). Then a contigs database was generated for each *Methanohalophilus* genome using the anvi-gen-contigs-database function in Anvi'o using the --external-gene-calls flag to import in-house gene calls generated from MetaProdigal (120). Next using the corresponding in-house annotations, the anvi-import-functions was used to import in-house annotations.

The overall pangenome was calculated using the anvi-pan-genome blast and an mcl-inflation of 10, due to the inclusion of MAGs. Bins of gene clusters unique to subsurface and

surface genomes were obtained using the Anvi'o interactive software. As described previously (125), this pangenomic workflow calculates similarities of each amino acid sequence in every genome against every other amino acid sequence using blastp (126), then weak hits are removed using the 'minbit heuristic' (here we used the default of 0.5) (127), and gene clusters are identified using the MCL algorithm (128). Next the number of occurrences of each gene cluster per genome and the total number of gene clusters in each genome contains was calculated. Hierarchal clustering of gene clusters based on their distribution across genomes and genomes based on gene clusters they share was calculated using Euclidean distance and Ward clustering. Average Nucleotide Identity (ANI) among genomes (Figure 2.5, 2.6) was calculated using the anvi-compute-ani function in Anvi'o (124).

Phylogenetic trees of *Methanohalophilus* single copy core amino acid sequences and 63 concatenated ribosomal amino acid sequences (91) were generated using Protpipeliner, a python script developed in-house for generation of phylogenetic trees (https://github.com/TheWrightonLab). Briefly, a maximum likelihood phylogeny for each muscle alignment was conducted using RAxML version 8.3.1 under the LG+α+γ model of evolution with 100 bootstrap replicates. All phylogenetic trees were visualized in iTOL.

2.4.7 Viral analyses and CRISPR arrays

The viral genome shown in Figure 2.9 that linked to 61 *Methanohalophilus* spacers was previously reported in (15). The viral genome was recovered from a Marcellus metagenome (Marcellus 1 from day 328 post hydraulic fracturing as reported in (44)), NCBI accession number SAMN04417546.

The CRISPR Recognition Tool plugin (CRT, version 1.2) in Geneious was used to identify CRISPR arrays in *Methanohalophilus* isolate genomes and MAGs. To identify matches

between viral protospacers and *Methanohalophilus* CRISPR-Cas array spacers (as well as comparing *Methanohalophilus* spacers among genomes) we used BLASTn with an *E*-value cutoff of $1e^{-5}$. All matches were manually confirmed by aligning sequences in Geneious; one bp mismatch was allowed. Matching spacers among *Methanohalophilus* genomes are included in Figure 2.9 and Appendix A. It should be noted that directionality of CRISPR arrays was not inferred. Links between viral sequences and *Methanohalophilus* were used to construct Figure 2.9. *Methanohalophilus* CRISPR-Cas systems were classified by manually examining the CRISPR-Cas proteins of annotated contigs (114).

2.4.8 Growth rate experiment

Methanohalophilus WG1-DM and *Methanohalophilus* WG1-MB isolate cultures were grown on modified DSMZ 479 media (see above) with trimethylamine, dimethylamine, monomethylamine, and methanol as a separate carbon sources with a concentration of 5mM. Prior to inoculation, cells were washed anoxically using a no carbon substrate modified DSMZ 479 media and inoculated to the same OD_{600} (approximately 0.15) in high salt media. A portion of the cells were boiled in water for 30 minutes and inoculated into the same media and substrates for a control. Growth curves were done in triplicate at 37°C for each treatment with one substrate addition at the time of inoculation. Using optical density measurements at 600nm as an analog for microbial growth, turbidity was measured over a period of 8.5 day (206 hours). Isolate methane production was quantified at beginning and end time points using a Shimadzu (GC-2014) gas chromatograph equipped with a thermal conductivity detector (TCD) using helium as a carrier gas at 100°C.

Table 2.1 Microbial studies of hydraulically fractured shale produced fluids.

STUDY	ARCHAEA ANALYZED	*Methanohalophilus*	LOCATION	FORMATION (# OF WELLS)	METHOD
Borton, *et al.* (2018)	Yes	Yes	Ohio and West Virginia	Marcellus (3) and Utica (2)	Metagenomics
Daly, *et al.* (2016)	Yes	Yes	Pennsylvania	Marcellus (1)	Metagenomics
Lipus, *et al.* (2016)*	Yes	Yes	Unknown	Marcellus (1)	Metagenomics
Tucker, *et al.* (2015)	Yes	Yes	Pennsylvania	Marcellus (4)	Metagenomics
Akob, *et al.* (2015)	Yes	Yes	Pennsylvania	Marcellus	Enrichment 16S rRNA gene analysis
Cluff, *et al.* (2014)	Yes	Yes	Pennsylvania	Marcellus (3)	16S rRNA gene analysis
Davis, *et al.* (2012)	No	N/A	Texas	Barnett (2)	16S rRNA gene analysis
Fichter, *et al.* (2012)	Yes	Yes	Unknown	Haynesville (1)	16S rRNA gene analysis
Liang, *et al.* (2016)	Yes	Yes	Texas	Barnett (6)	16S rRNA gene analysis
Mohan, *et al.* (2013) a	No	N/A	Pennsylvania	Marcellus (1)	16S rRNA gene analysis
Mohan, *et al.* (2013) b	Yes	Yes	Unknown	Marcellus (1)	Impoundment 16S rRNA gene analysis
Struchtenmeyer, *et al.* (2012)	No	N/A	Texas	Barnett (7)	16S rRNA gene analysis
Waldron, *et al.* (2007)	Yes	Yes	Michigan	Antrim (8)	Enrichment 16S rRNA gene analysis
Wuchter, *et al.* (2013)	Yes	Yes	Michigan	Antrim (3)	Enrichment 16S rRNA gene analysis

*Reads were not deposited for these metagenomic studies, thus relative abundance could not be assessed.

Table 2.2 Overview of 11 *Methanohalophilus* genomes and their designated species and/or strain numbers used in this study.

GENOME	SIZE (MB)	TYPE	SOURCE	ACCESSION	REFERENCE
Methanohalophilus euhalobius DSMZ 10369	1.87	Isolate	Oil field *Russia*	SAMN06296050	this study
Methanohalophilus euhalobius WG-1MB	1.98	Isolate	HF shale *Ohio, USA*	SAMN06295989	this study
Methanohalophilus sp. 3-U2	1.89	MAG	HF shale *Ohio, USA*	SAMN06267276	this study
Methanohalophilus sp. 4-M4	1.60	MAG	HF shale *West Virginia, USA*	SAMN06264872	this study
Methanohalophilus sp. WG1-DM	1.94	Isolate	HF shale *Ohio, USA*	SAMN07462262	this study
Methanohalophilus halophilus Z-7982	2.03	Isolate	Microbial Mat *Shark Bay, Australia*	CP017921	L'Haridon, et al. 2017
Methanohalophilus mahii SLP DSM 5219	2.01	Isolate	Great Salt Lake *Utah, USA*	CP001994.1	Spring, et al. 2017
Methanohalophilus portucalensis FDF-1	2.09	Isolate	Salt Pan *Portugal*	CP017881.1	L'Haridon, et al. 2018
Methanohalophilus sp. DAL1	1.89	MAG	HF shale *Pennsylvania, USA*	SAMN05258748	Lipus, et al. 2017
Methanohalophilus sp. 2-GBENRICH	1.89	MAG	HF shale *Ohio, USA*	SAMN05172267	Borton, et al. 2018
Methanohalophilus sp. T328-1	2.08	MAG	HF shale *Pennsylvania, USA*	SAMN04432769	Daly, et al. 2016

Table 2.3 Range and optimum salinities for isolated *Methanohalophilus spp.*

Methanohalophilus spp.	OPTIMUM CHLORIDE CONCENTRATION (G/L)	RANGE OF CHLORIDE CONCENTRATION (G/L)	REFERENCE
Methanohalophilus. portucalensis FDF-1	77	17-123	Boone, et al. (1993)
Methanohalophilus. mahii SLP	71	N/A	Paterek, et al. (1988)
Methanohalophilus levihalophilus DSM 2094	13	7-46	Katayama, et al. (2014)
Methanohalophilus strain Z7302	88	60-155	Lai, et al. (1992)

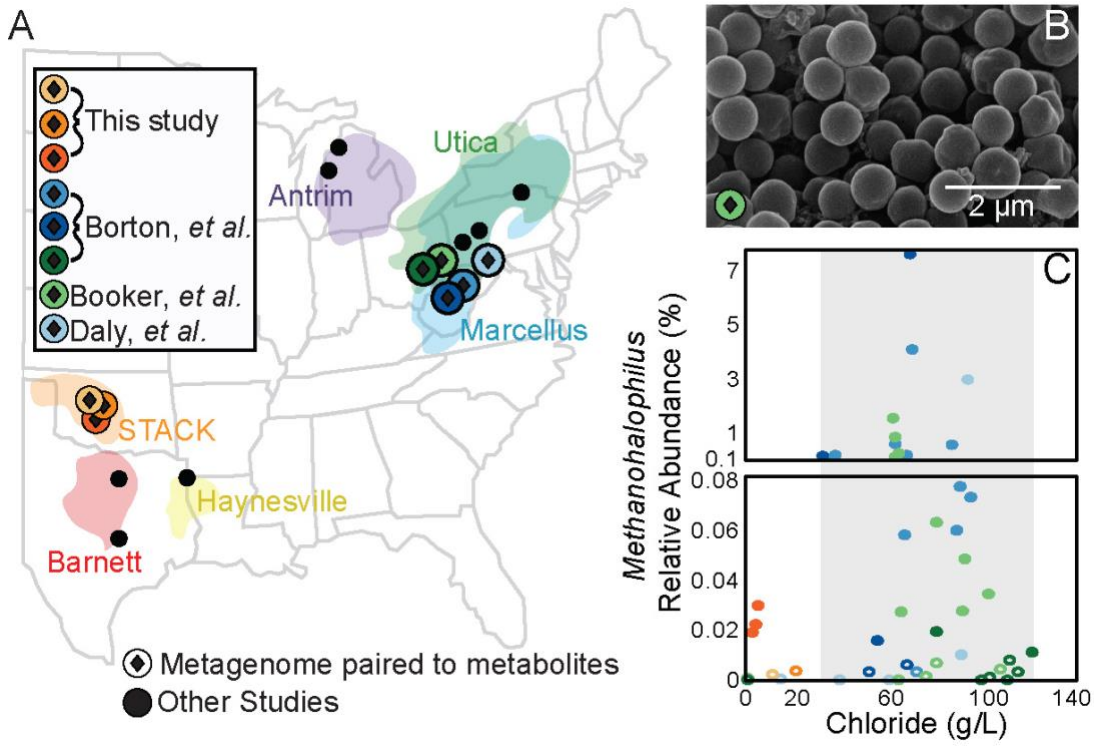

Figure 2.1 *Methanohalophilus* prevalence across shales and salinities.

(A) Initial work using metagenomics and single marker gene analyses to identify microorganisms from 44 natural-gas wells distributed across several geographically distinct unconventional reservoir formations (Antrim, Utica, Marcellus, Barnett, Haynesville, and STACK). Only studies that surveyed archaeal diversity are shown. Previously published studies denoted by a colored circle with a diamond have time series metagenomic data with paired geochemical and metabolite data. Filled black circles represent studies that used 16S rRNA data to determine the presence of *Methanohalophilus* (B) SEM image of *Methanohalophilus* WG1-MB isolated from flowback and produced fluids from the Utica/Point Pleasant Formation. (C) *Methanohalophilus* genome relative abundance across 5 natural-gas wells with metagenomic data graphed with chloride concentrations shows the salinity window (grey box) for *Methanohalophilus* occurrence in hydraulically fractured unconventional reservoir wells. Color of dot denotes natural-gas well and corresponds to map (A), while open circles denote absence of *Methanohalophilus* reads or scaffolds >2 kb (see methods).

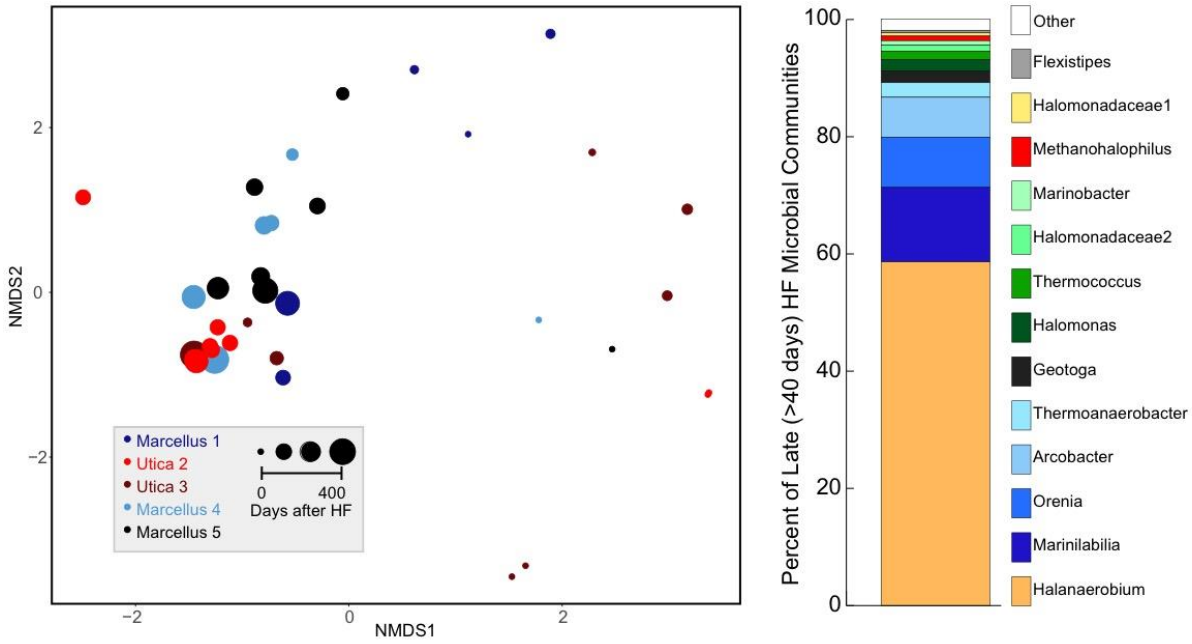

Figure 2.2 Non-metric multidimensional scaling shows that microbial communities converge at late timepoints.

Samples are colored by well and bubble size denotes time after HF. Bar graph denotes the percent membership of late (more than 40 days post hydraulic fracturing) microbial communities. Notably, *Methanohalophilus*, is the 11th most reoccurring member at late (>40 after hydraulic fracturing) time points.

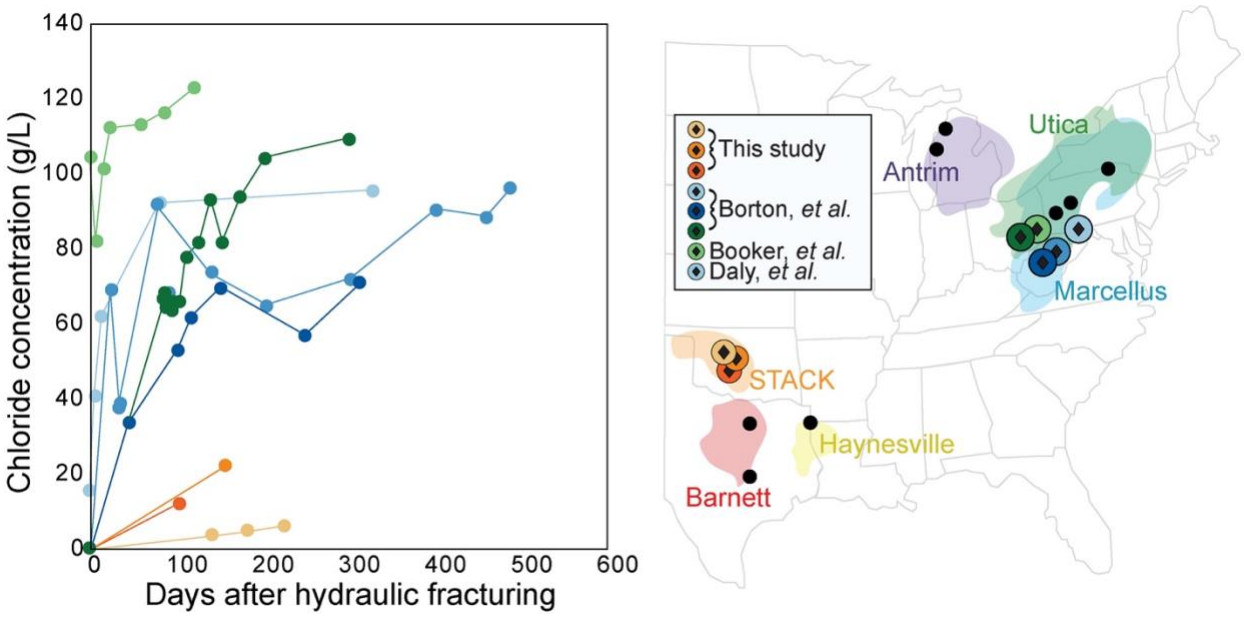

Figure 2.3 Salinity increased overtime after hydraulic fracturing.

Chloride concentrations over time for field produced and input fluids are shown, with color denoting well and corresponding to Figure 2.1A (map).

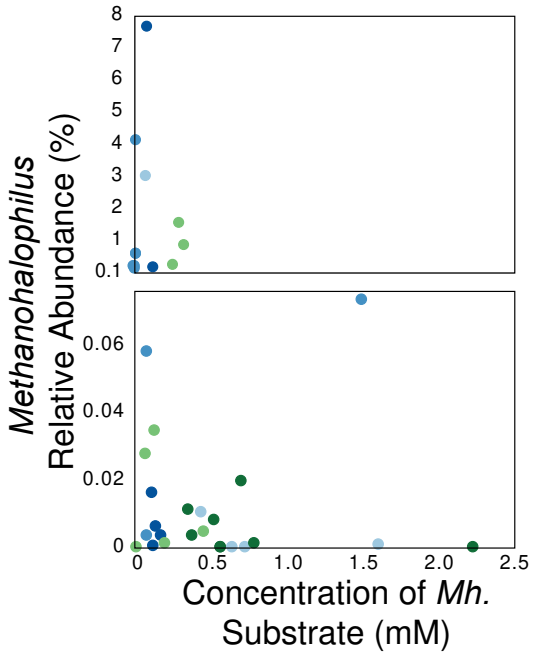

Figure 2.4 *Methanohalophilus* abundance correlated to salinity.

Methanohalophilus genome relative abundance across 5 wells with metagenomic data graphed with possible *Methanohalophilus* substrate (trimethylamine, dimethylamine, monomethylamine, and methanol) concentrations. Color of dot denotes well (Utica in green and Marcellus in blue, corresponding to Figure 2.1A).

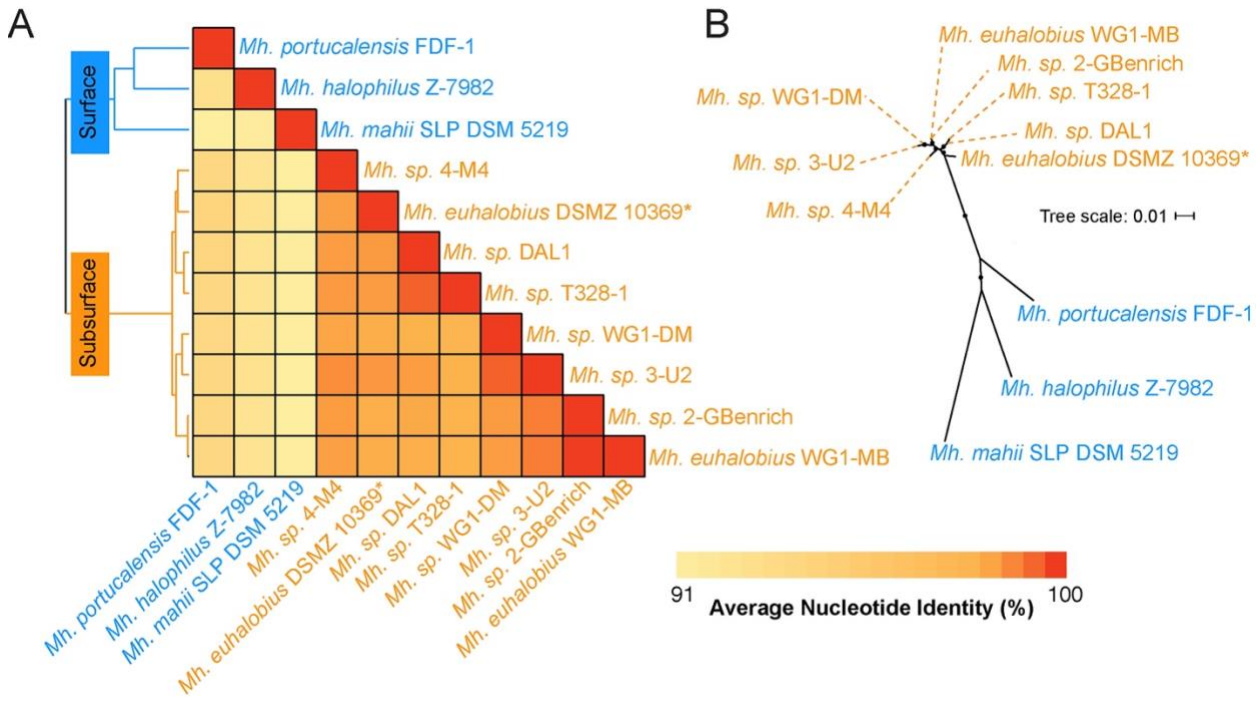

Figure 2.5 Subsurface derived genomes have a different genomic signature compared to surface derived genomes.

(A) Heatmap shows pairwise comparisons of average nucleotide identities (ANI) across 11 *Methanohalophilus* genomes and (B) a maximum likelihood tree of 682 core genes found across 11 *Methanohalophilus* genomes. Closed circles on phylogenetic tree are located at every branch point and represent bootstraps >99. All subsurface genomes were from hydraulically fractured unconventional reservoir except one, denoted by an asterisk (*).

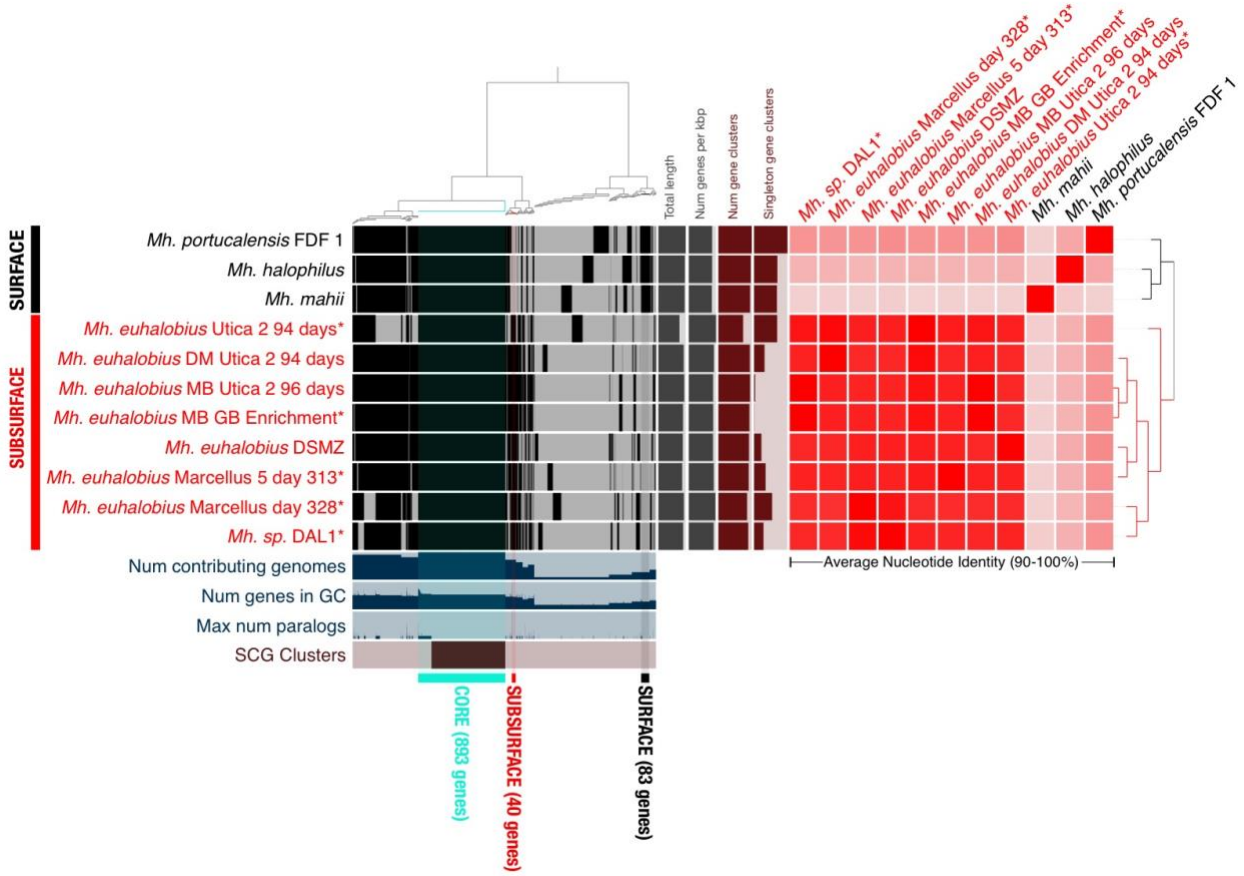

Figure 2.6 Comparative genomics of surface and subsurface derived *Methanohalophilu*s.

Anvi'o (124) pangenome display of 11 *Methanohalophilus* genomes, with MAGs marked with an asterisk (*). Subsurface (red) versus surface (black) genomes are indicated on the left, with each row representing 1 genome across the entire figure. Dendrogram at the top represents the hierarchical clustering of gene clusters based on occurrence within each genome, with each vertical line representing 1 of 1,016 gene clusters. Gene clusters are grouped by Core, Subsurface, or Surface at the bottom. Heatmap on the right shows and all-versus-all comparison of the Average Nucleotide Identity (ANI) of all genomes shown, with the dendrogram on the left denoting the hierarchal clustering of genomes based on ANI.

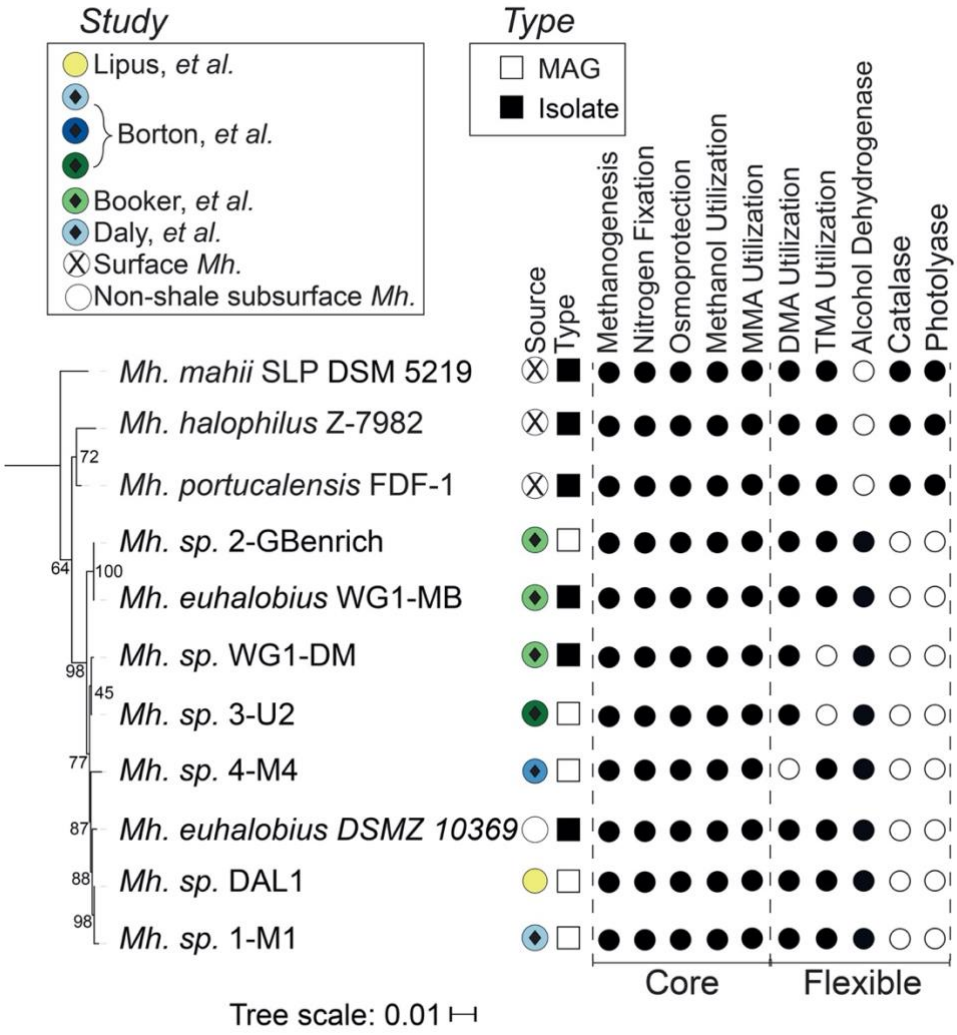

Figure 2.7 Phylogenetic tree of 63 concatenated ribosomal proteins from *Methanohalophilus* and the genetic traits unique or shared across this genus.

Presence and absence of selected genes in the core and flexible genome (Appendix A) are denoted by bubble shading (black, presence; white, absence). The source of genome denoted in the first column is based on bubble fill, with surface (white bubble with an X), conventional oil well (white bubble) and unconventional reservoir play (colored bubbles from Figure 2.1A).

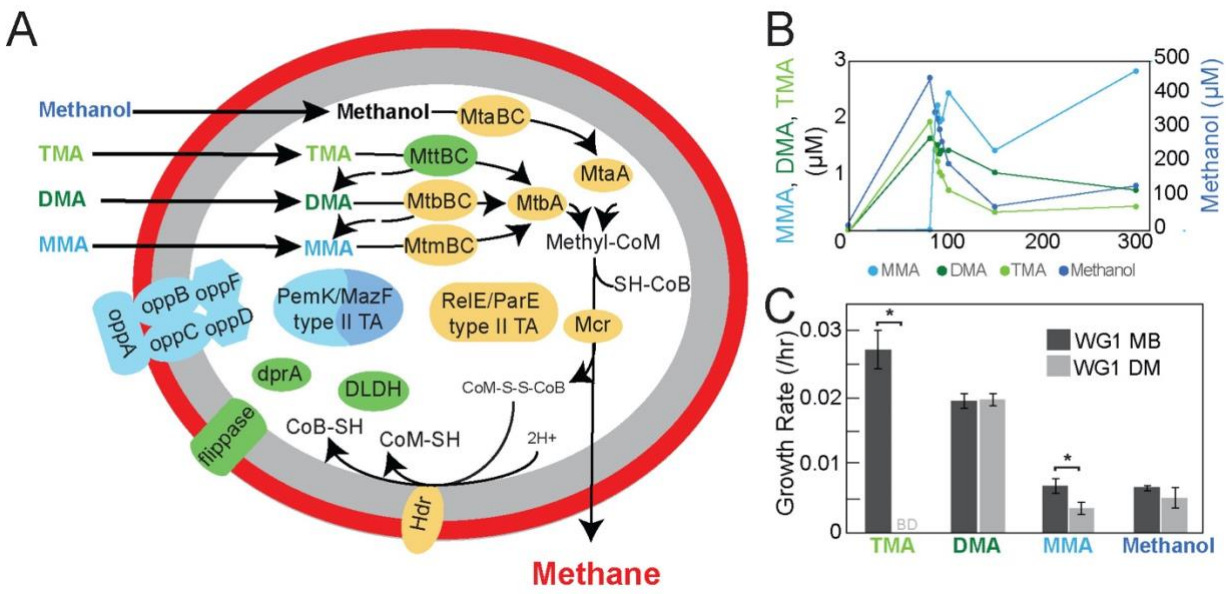

Figure 2.8 Genomic and physiological comparison of *Methanohalophilus euhalobius* strains.

Two *Methanohalophilus* strains, *Methanohalophilus euhalobius* WG1-DM and *Methanohalophilus euhalobius* WG1-MB, were isolated from produced fluids originating from the same Utica/Point Pleasant Formation natural-gas well two days apart. These isolates are indistinguishable by 16S rRNA sequence, yet their genomes harbor ecologically relevant differences (A). Gene symbols are colored based on their genomic presence (yellow, present in both genomes; green, present only in *Methanohalophilus euhalobius* WG1-MB; blue, found only in *Methanohalophilus euhalobius* WG1 DM). (B) *Methanohalophilus* substrates are present in flowback and produced fluids. (C) Genomic inferences are validated between the two strains, as the growth rates of strains WG1-DM and WG1-MB are different based on substrate. Error bars represent one standard deviation.

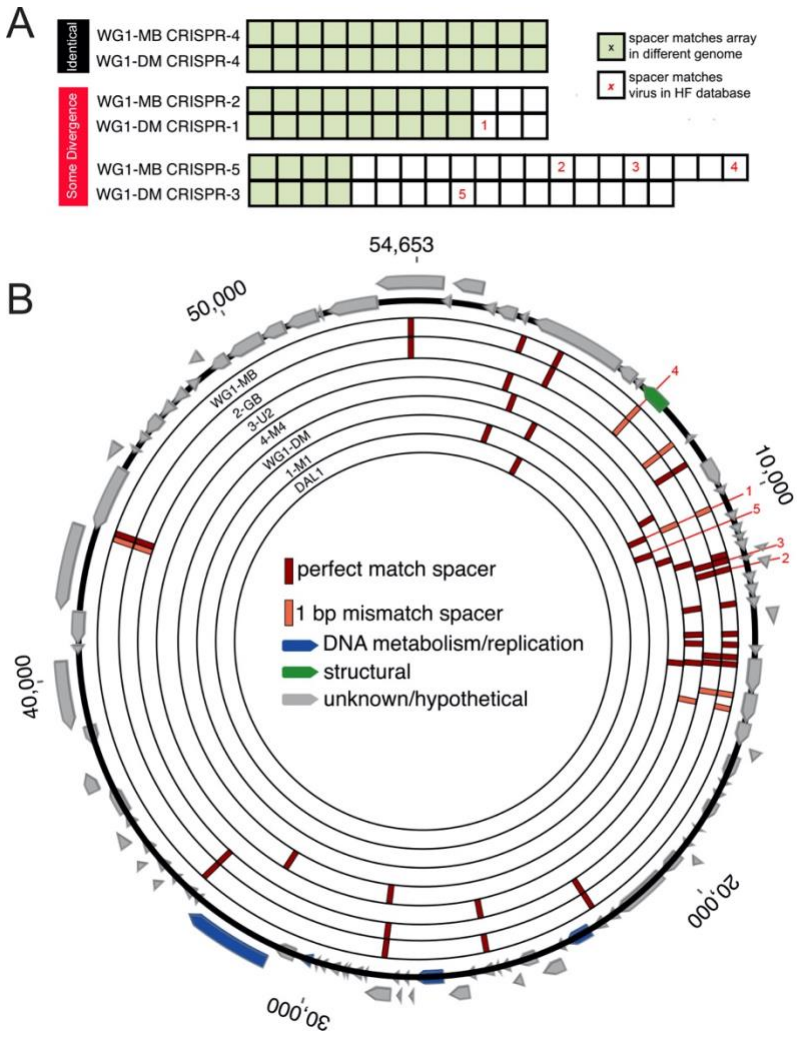

Figure 2.9 *Methanohalophilus* CRISPR-Cas array comparisons.

(A) Sequence comparisons of CRISPR-Cas arrays from two isolates from the same natural-gas well highlight *Methanohalophilus* strain level microdiversity, with spacer sequences in an array denoted by a box (white boxes are unique to that genome and green boxes are identical matches shared between genomes). Pairwise comparisons between the seven other *Methanohalophilus* genomes containing CRISPR-Cas arrays can be found in Appendix A. (B) 61 *Methanohalophilus* spacers hit to the same viral population, represented here by a 54,653 bp circular viral genome (virus M1_T328_scaffold_10). Predicted viral ORFs are shown and colored according to their annotation in the legend. From the 7 *Methanohalophilus* genomes that contain a spacer match to this virus, we show the unique viral genomic positions targeted by the spacers. The viral genome position is denoted by red/orange bar (colored depending on mismatches), and each *Methanohalophilus* is represented by a separate concentric circle. Red numbers in (A) identify the 5 spacers from WG1-MB and WG1-DM genomes that target (match) portions of the viral genome shown by the same red numbers in (B).

Chapter 3: Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales[2]

3.1 Introduction

In 2016, natural gas became the main source of electricity in the United States—the first time in history that a natural resource other than coal has provided a bulk of the nation's power (129). Sixty percent of the natural gas produced in the United States comes from hydraulically fractured shales, a majority of which is generated in Ohio, West Virginia, and Pennsylvania (129). Hydraulic fracturing (HF) is the high-pressure injection of water, chemical additives, and proppant into the earth's subsurface to fracture hydrocarbon-bearing shales, thereby releasing economically important trapped natural gases. This process unintentionally creates a new microbial ecosystem, where a subset of surface-derived microorganisms proliferate in shales more than 2,500 meters below the earth's surface (15).

Recent research suggests that microbial life in shales may impact gas and oil production efficiencies (59, 66). For instance, the persistence of methanogens in these ecosystems may contribute to increased biogenic methane formation by *Methanohalophilus*, while negative impacts, such as corrosion and sulfidogenesis ('souring'), are associated with other prevalent microbial community members including *Halanaerobium* (15, 59, 62–64, 66, 130, 131). To grow in fractured shales, microorganisms must adapt to increased salinities and reduced chemical conditions where fermentative metabolisms prevail (59). Given these

[2] This chapter was reproduced verbatim from "Borton, et al. Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. *PNAS* (2018)". 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.

selective pressures, persisting shale-hosted microbial communities are constrained to several halotolerant members, including *Halanaerobium* and *Methanohalophilus* which co-occur across every fractured shale sampled to date (59). Metagenomic and metabolite analyses from a single well suggested that glycine betaine, an amino acid derivative, may play an important role as an osmoprotectant and as an energy source for these co-occurring shale organisms (15). However, the glycine betaine supported metabolisms employed across geographically and geologically distinct fractured shales remains unknown.

Here, we use a combination of field investigations and detailed laboratory microcosm experiments to define the metabolic network supported by glycine betaine. First, we sample glycine betaine prevalence and concentration in the field using temporally collected fluid samples collected from Utica and Marcellus fractured shale wells. We then established laboratory microcosm reactors with Utica produced fluids collected 96 days after HF and used proteomics to define the impacts of glycine betaine on persisting shale microbial communities. To extend these laboratory-discovered processes back to the field scale, we conducted the first paired metagenome and metabolome analysis from over 40 samples collected across five fractured shale wells located in the Appalachian basin. This comprehensive dataset offers unique insight into previously cryptic amino acid based metabolisms that may sustain life in these economically important ecosystems.

3.2 Results and Discussion

3.2.1 From the field to the lab: Constructing model shale microbial communities in the laboratory

To understand the broader importance of glycine betaine across geographically distinct fractured shales, we profiled glycine betaine concentrations in input (fluids injected during HF)

and produced fluids from five shale wells sampled up to 600 days after HF (Figure 3.1, Appendix B-C). Glycine betaine was present in all hydraulically fractured shale wells, with two of the five wells showing a trend where glycine betaine is not detected in the input fluids but is then produced and maintained *in situ* (Marcellus 1 and Utica 2). For these two wells, glycine betaine was positively correlated to salinity (Pearson, $R= 0.87$, $p<0.001$), corroborating our prior hypothesis that this metabolite is likely microbially synthesized *in situ* to support microbial adaptation to brine level salinities (15). In the other three wells, glycine betaine was detected in the input fluids, albeit at low concentrations (>0.8 μM). This could be a result of operators using recycled produced fluids as input fluids or the exogenous addition of glycine betaine as a surfactant amended to input fluids ((132), https://fracfocus.org/chemical-use). Glycine betaine dynamics in these wells hint at both microbial utilization and production, however it is also possible that glycine betaine is leached from the dissolution of shale rock (Figure 3.1).

To understand the possible sources and metabolic roles of this prevalent metabolite, we generated laboratory microcosms using produced fluids collected 96 days post HF (Utica well 2, Figure 3.1, red arrow). To identify the microbial sources of glycine betaine, these reactors were established without shale rock. Triplicate anoxic microcosms were amended with and without glycine betaine in a chemically undefined medium containing yeast extract (see Methods for recipe) and incubated for 20 days, with three time points chosen for metabolite, metagenome, and metaproteome analyses. Abiotic controls showed no metabolite changes through the experiment (Figure 3.2). Time points were collected at the beginning (T_0), at maximum cell density on day 2 (T_M), and upon substantial methane production (1.5 log fold increase from T_0) on day 20 (T_F) (Appendix C). Metagenomic sequencing facilitated the reconstruction of four draft genomes belonging to the genera *Halanaerobium*, *Methanohalophilus*, *Geotoga,*

and a novel genus within the Clostridiales (Figure 3.3-3.4, Appendix C). The organisms from which these genomes were reconstructed were the only members of the microbial community in both the glycine betaine and non- glycine betaine enrichment cultures at all time points (Figure 3.5). This enrichment reflects the low genus-level diversity previously reported in late produced fluids from Utica and Marcellus shales (15, 59, 133).

Genomes reported here were estimated to be greater than 93% complete, with less than 2% contamination, and contained full-length 16S rRNA genes (Appendix C). Based on the recently proposed Genomic Standards Consortium standards (134), the genomes recovered here would be considered high-quality. The unassigned Clostridiales genome is most closely related to *Dethiosulfatibacter* by 16S rRNA gene analysis (~90% identity, SILVA) and *Dethiosulfatibacter aminovorans* by average nucleotide identity at the genome level (73.1%) (Figure 3.3-3.4). Following the naming convention for genomes assembled from metagenomes (135), we propose the genus name *Candidatus* Uticabacter based on the shale formation from which this genome was recovered. 16S rRNA gene fragments (V4 region) were identical to the near-complete 16S rRNA gene recovered from our *Candidatus* Uticabacter genome, suggesting that members of this genus have been previously detected in a hydraulically fractured shale well in the Sichuan Basin in China (NCBI SRR2094439.12567.1) (131). Beyond *Ca.* Uticabacter, the other members recovered in our laboratory genomic analyses are routinely reported in studies from fractured shales across the United States (74, 75, 79). For instance, 16S rRNA genes corresponding to *Halanaerobium* and *Methanohalophilus* co-occur in all but one of these 17 studies (15). Together these findings demonstrate that the microorganisms detected in our microcosms, and likely their metabolic interactions, are relevant to fractured shale ecosystems.

Next, we used metagenome-resolved metaproteomics to uncover the active metabolisms assigned to each genus. A total of 555,973 unique peptides were recovered from 15 metaproteomes, with an average of 37,046 unique peptides per microcosm sample (Appendix B-C). Across all time points and treatments, a majority of the proteins analyzed were from the genus *Halanaerobium* (63%). Proteins from other members of the microbial community were also detected, with 15% of total proteins from *Methanohalophilus*, 11% from *Ca.* Uticabacter, and 7% from *Geotoga* (Figure 3.1). Interestingly, overall protein content and taxonomic assignment could not be statistically differentiated between the glycine betaine and non-glycine betaine treatment at the middle time point, largely driven by the dominance and conserved metabolism of *Halanaerobium* across the two treatments. The final time point (T_F) was statistically differentiated by treatment, with proteins from *Methanohalophilus* enriched in the glycine betaine microcosm where methane was produced in high amounts, while proteins assigned to *Ca.* Uticabacter and *Geotoga* were more enriched in the non-glycine betaine microcosm that produced significantly less methane.

3.2.2 Osmoprotection mechanisms enabling salinity adaptation in laboratory reactors

Given the hypersaline conditions observed in late (>40 days post HF) produced fluids (Appendix B-C), we profiled microcosm metaproteomic data for evidence of osmoprotection strategies. While it has been well documented by our group, and others, that these organisms encode versatile osmoprotection strategies (15, 20, 71), the preferred mechanisms and how they change with extracellular availability of an osmoprotectant was unknown. Consistent with production and consumption patterns of glycine betaine across wells (Figure 3.1), all organisms in the microcosm have the potential to uptake glycine betaine, with *Methanohalophilus* exclusively utilizing the compatible solute strategy through uptake and

synthesis (Appendix B, Figure 3.6). From glycine, *Methanohalophilus* can produce glycine betaine through sarcosine and N-N-dimethylglycine intermediates (Appendix B). Notably, this pathway is expressed regardless of glycine betaine amendment, signifying that glycine betaine may be key to biogenic methane production in fractured shales.

Collectively, metaproteomic data indicate that *Ca.* Uticabacter and *Halanaerobium* likely use the salt-in strategy through sodium/proton antiporters, while *Methanohalophilus* and *Geotoga* are reliant on the osmolyte strategy (Appendix B, Figure 3.6). Inferring osmoprotectant function from meta-omics is complicated by the fact that many transporters are non-specific and often these compounds can play other roles in cellular assimilation or energy production. Despite these challenges, our findings expand upon prior reports that *Halanaerobium* solely uses a salt-in strategy and provides proteomic evidence for the use of choline uptake for osmoprotection (Appendix B, Figure 3.6). Given that glycine betaine has multiple assimilatory (osmoprotection, nitrogen and carbon source) and dissimilatory (energy generation) uses and is prevalent in fractured shales (Figure 3.1), we suggest glycine betaine may be a keystone metabolite. Here we used glycine betaine and non-glycine betaine amended laboratory microcosms to investigate the ecological interactions, including predation, mutualism, and competition, present in fractured shale microbial communities.

3.2.3 Viral predation and resistance is ongoing in laboratory reactors

To elucidate predator-prey interactions in these microcosms we identified viral genomes, linked these viruses to hosts, and measured their activity. Fifty-four assembled viral contigs were recovered and clustered into 16 unique populations, 25% of which were affiliated with the Order Caudovirales, while the remaining majority (75%) were taxonomically novel. Viral dynamics were coordinated to their host, and notably not impacted by glycine betaine amendment (Figure

3.7, Appendix B). Two of the viral populations found in this microcosm were also previously reported (15) in Marcellus well 1 (Figure 3.1, Appendix B-C). This finding demonstrates the relevance of these laboratory enriched viruses to the shale ecosystem.

We detected 326 unique viral peptides from 13 of the 16 viral populations (Appendix B-C). Most of the viral peptides were identified as proteins with unknown function (36%), however, peptides involved in virion production (e.g. terminase and head proteins) and viral integration into host genomes (e.g. resolvase and recombinase) were also detected (Figure 3.7, Appendix B). This expression data show that a majority of the microcosm viruses are active, and these include both temperate and lytic infections. Thus, fractured shale microbial communities are likely evolving under strong constraints exerted by a diverse set of viruses.

Previously we detected spacer incorporation in a *Halanaerobium* genome over time from field produced fluids (Marcellus Well 1) (15). Here we provide the first evidence for the activity of the CRIPSR-Cas system from deep biosphere microbial communities. Cas proteins for all three functional stages of adaptive immunity were expressed (adaptation, expression, interference, Figure 3.8) (136). Of particular importance, both *Methanohalophilus* and *Halanaerobium* expressed adaptive proteins for incorporating spacers into CRISPR loci (Cas1), as well as interference proteins for producing cognate RNAs (Cas5) that bind to and cleave the viral DNA (Cas 3, *Methanohalophilus* only) (Appendix B). The congruence between laboratory and field viral populations and evidence of CRISPR-Cas activity demonstrate that the strong viral predation captured in our laboratory microcosms reflects ongoing viral-host interactions maintained at the ecosystem scale.

3.2.4 Mutualistic interactions sustain biogenic methane production in laboratory reactors

Consistent with our prior metagenome findings and physiological characterizations of the genus (15, 86, 137), *Methanohalophilus* is inferred to be an obligate methylotrophic methanogen, lacking the capacity to utilize hydrogen or acetate. Additionally, this genome lacked the genes necessary to directly use quaternary amines like choline and glycine betaine (27, 28, 138). *Halanaerobium* appears to be an obligate fermenter, as the genome lacks an electron transport chain and terminal oxidase or reductase genes (64). We have previously suggested based solely on metagenomic inferences that the fermentation of the amino acid derivative glycine betaine will yield products sustaining methylotrophic methanogens in fractured shales (15).

To better elucidate this metabolic cross-feeding, we used linear discriminant analysis to identify and report the significant metabolisms occurring at different stages of biogenic methane production (LEfSe (139), Appendix B-C). Our proteomics data revealed that glycine betaine was fermented by *Halanaerobium* to yield trimethylamine (TMA) at the middle time point, which sustained methanogenesis at the later time point. The proteins necessary for this metabolic symbiosis (*Halanaerobium* GrdHI and *Methanohalophilu*s MttB) were discriminating features of the middle and final time points respectively, and we failed to identify any other sources for TMA production (Figure 3.9). Other possible sources of methane include methanol (MtaB), monomethylamine (MtmB), and dimethylamine (MtbB) but not acetate, as corresponding proteins were detected for methylotrophic substrates only. Our findings are consistent with prior reports where methylotrophic methanogenesis is more prevalent in saline ecosystems, likely because this methanogenesis pathway (rather than hydrogenotrophic or acetoclastic alternatives) generates higher energy yields that are needed to sustain the increased cost of osmoprotectant synthesis (21, 59).

Metabolite analysis supported the metaproteomic results, revealing that 90% of glycine betaine consumed in the first two days was recovered as TMA. 95% of this TMA was subsequently converted to methane by the last time point. Interestingly in the non- glycine betaine reactors, the proteomic and metabolomic patterns are similar but less prominent, with 60% of *Halanaerobium* produced TMA converted to methane. The fact that glycine betaine metabolism occurs regardless of experimental manipulations has ramifications for *in situ* processes, as the substrate concentrations in the non-glycine betaine microcosm were similar to field conditions (Figure 3.1). Moreover, the synthesis of glycine betaine in the shale-free laboratory microcosm supports our supposition that increased glycine betaine over time in the field derived produced fluids was due to microbial synthesis (Figure 3.1). The ubiquity of *Methanohalophilus* across fractured shales (59) and the high efficiency of methane production demonstrated here indicate that methylamine methanogenesis may be active and important to shale natural gas production. Supporting our findings, a prior study predicted that biogenic methane accounted for 12% of methane produced in a shale-gas well lifetime (66). Our findings leave open the possibility that the augmentation of fractured shales with exogenous methyl-C1 compounds could enhance biogenic methane production down well, analogous to acetate amendment techniques currently employed in coal-bed methane recovery (68).

We next examined the capacity for other Stickland fermentations that support methanogenesis. Similar to *Halanaerobium*, *Ca.* Uticabacter expressed proteins to ferment sarcosine (sarcosine reductase, GrdFG) (35), yielding monomethylamine that *Methanohalophilus* utilizes for methane production (Figure 3.10-3.11, Appendix C). Monomethylamine concentrations and necessary enzymes (MtmBC) followed the same pattern as trimethylamine but were significantly lower (Figures 3.2, 3.10, Appendix B). Unlike glycine betaine, sarcosine

does not decrease with monomethylamine formation, but rather increases over time in both biological treatments, suggesting microbial sarcosine production exceeds its removal (Appendix B). We show that mutualistic exchange of methylamines produces biogenic methane in fractured shale microbial communities.

3.2.5 Untangling the Stickland fermentation network revealed substrate partitioning and competition in laboratory reactors

While our field and laboratory studies indicated that glycine betaine is readily reduced to TMA by the prevalent and highly dominant shale bacterium *Halanaerobium* (59, 130), the amino acid electron donor for this fermentation was unknown. Our laboratory study illuminated the genomic potential for utilizing known Stickland electron donors and acceptors in a shale-derived microbial community, with the reactions and key functional genes for these metabolisms summarized in Table 3.1.

Based on coupled meta-omic data from the glycine betaine enrichment, we conclude that lysine is likely the primary electron donor used by *Halanaerobium* to reduce glycine betaine to TMA (Figure 3.10). Using the enzyme 3,5-diaminohexanoate dehydrogenase, *Halanaerobium* is the only bacterium to oxidize lysine to acetate, butyrate and ammonia through crotonyl-CoA in the microcosm (Appendix B). The pattern of expression for this enzyme was significantly correlated to that of glycine betaine reductase ($p<0.01$), and metabolite stoichiometry demonstrated that 93% of the lysine was oxidized in the first two days during primary glycine betaine reduction. Of the other possible Stickland electron donors (24, 37, 140), lysine was in the greatest concentration, accounting for up to 17% of glycine betaine reduction, while other *Halanaerobium* Stickland donors implicated by proteomics and metabolomics included serine (7.2%), methionine (6.7%), glycine (4.1%) and threonine (3.8%) (Appendix B-

C). Given that only a little more than a third of glycine betaine reduction can be accounted for from known amino acid reductants in the Stickland reaction, *Halanaerobium* also uses hydrogen or other currently unknown reductants as the electron donor for glycine betaine reduction.

Unlike lysine, which represents a non-competitive substrate for *Halanaerobium*, other Stickland electron donors are more widely used by members of the community. In addition to *Halanaerobium*, *Ca.* Uticabacter can also compete for glycine as a Stickland electron donor to support sarcosine reduction, or may use glycine as both a donor and acceptor simultaneously (23, 35, 36) (Appendix B). Glycine is consumed at all time points except at the last time point of the no-glycine betaine amendment (Figure 3.10). From proteomic and metabolite analysis, we infer that *Geotoga* is responsible for this glycine production, via operation of the glycine cleavage system in reverse (Figure 3.12, Appendix B), using ethylene glycol as an oxidant. Overall, glycine is the most interconnected metabolite based on its variety of uses in the microcosm community (Figure 3.12).

In summary, the laboratory microcosms demonstrated that glycine betaine and glycine have both adaptive and metabolic roles in fractured shale communities. For instance, glycine betaine is synthesized and used as an osmoprotectant by *Methanohalophilus*, while *Halanaerobium* utilizes glycine betaine to produce energy, providing *Methanohalophilus* with substrates. Similarly, *Methanohalophilus* uses glycine to synthesize glycine betaine for osmoprotection, while *Ca.* Uticabacter uses glycine to reduce sarcosine to the methanogenic substrate, monomethylamine. In addition to amino acids, sugars like trehalose and maltose can also be used as an energy source (*Halanaerobium*) and an osmoprotectant (*Geotoga*). Overall our study focuses on the multiple uses for amino acids (and their derivatives) in facilitating microorganism growth and maintenance in up to 2,500-meter deep fractured

shales. Also hinting at the importance of organic nitrogen to rock-hosted systems, Lloyd, *et al.* (141) demonstrated the significance of detrital proteins to supporting life in deep marine sediments. It is intriguing to speculate that these nitrogen transformations may be a conserved metabolism in the deep biosphere.

3.2.6 New shale metabolisms and end products discovered in laboratory reactors

In addition to predation, mutualism, and competition, we identified non-competitive substrates that provide energy for a single organism. Our proteomics data showed *Halanaerobium* uniquely fermented ethanolamine (EutBCEGH) and trehalose (TrePP) (Figure 3.2, 3.13), with the former substrate likely relevant to shale where ethanolamine is provided exogenously as a corrosion inhibitor and endogenously through biomass turnover of cell membranes (142). Collectively, the interconnected amino acid and sugar fermentations result in the buildup of methane, ammonium, formate, and acetate (Figure 3.11). Acetate was the most abundant produced metabolite, with the greatest production rate occurring before T_M (Figure 3.11). As expected, *Halanaerobium*-mediated glycine betaine reduction was responsible for the increased concentration of acetate between the two treatments, accounting for 97% of the difference between amended and non-amended glycine betaine microcosms. Congruent with the observed accumulation of acetate in microcosm studies, *Geotoga*, *Ca.* Uticabacter, and *Halanaerobium* all expressed genes for acetate production, with a 4-9 fold greater expression of acetate kinase from *Halanaerobium*. Other carbon sources supporting acetate production include trehalose and ethanolamine fermentation by *Halanaerobium* and ethylene glycol fermentation by *Halanaerobium* and *Geotoga*, which together could explain a third of the acetate produced in the non-amended reactors (Appendix B-C). The fermentation of ethylene glycol may be

important to fractured shales in the field, where this compound is commonly added to input fluids for use as a gelling agent in hydraulic fracturing (https://fracfocus.org/chemical-use).

3.2.7 Extending laboratory reactor findings to the field scale: microcosm generated hypotheses are validated in Appalachian Basin produced fluids

Batch-operated laboratory microcosms more readily permitted the quantification of gas and metabolic waste products generated by the shale microbial consortia (Figure 3.11), allowing mass-balance calculations that are currently not feasible at the field scale. Key outcomes from our laboratory microcosms included *(i)* deciphering tradeoffs between osmoprotectant and energy use, *(ii)* unveiling the pervasive Stickland fermentation network, and *(iii)* discovering new interconnected metabolites that may be essential to the shale metabolic economy (Figure 3.11). Specifically, we demonstrated that glycine betaine is a keystone metabolite that is not only vital to salinity adaptation, but also is fermented to TMA and acetate by *Halanaerobium*, a metabolism that subsequently fuels methane production by *Methanohalophilus*. Glycine was the most connected metabolite, with proteomics indicating use as an energy source for *Halanaerobium* Stickland reactions, transportation into the cell for osmoprotectant generation by *Methanohalophilus*, and intracellular synthesis for assimilatory purposes by *Geotoga*.

To quantify the relevance of these laboratory identified processes at the field scale, we analyzed input and produced fluid metagenomic and metabolite data. This includes samples from one previously published well (n=5, (15)) and 36 samples from four additional wells reported here. For each well, samples span input fracture fluids to produced fluids collected up to at least 300 days after HF. Along this time scale, fluids transition from freshwater to hypersaline (>35 chloride g/L) (Figure 3.14). From our field metagenome data, we defined microbial strain membership and relative abundance across the samples using a single copy,

conserved marker gene, RpsC (see Methods, Appendix B-C). Across these produced fluid metagenomes, we identified multiple strains of *Halanaerobium* and *Methanohalophilus* (9 and 3 strains, respectively) and a single strain of *Geotoga. Ca*. Uticabacter was removed from this analysis due to detection in less than 5 samples. For comparison, reliance on 16S rRNA gene diversity would have only resolved a single sequence type for each of these genera, showing this strain-level resolution better captured the genotypic microdiversity previously observed in shale fluids (15).

Consistent with the laboratory reactor data, metabolites related to osmoprotection were highly correlated in produced fluids across shale wells, with glycine betaine, choline, sarcosine, and creatine positively correlated to chloride (Figure 3.15, Appendix B). Of these compounds, glycine betaine is generally regarded as the most potent osmoprotectant (20), and thus it is possible that sarcosine and creatine may instead support glycine betaine biosynthesis as outlined in the Appendix B. Given that several of these metabolites are detected in the input fluids and are known additives in the fracturing process (https://fracfocus.org/chemical-use) (Figure 3.14), this finding provides further evidence that chemicals added during HF support life in this man-made ecosystem.

Like our laboratory microcosms, in the field, Stickland metabolites have significant coordinated associations. glycine betaine was positively correlated to TMA across produced fluids from fractured shales, supporting the notion that this osmoprotectant can be fermented to yield methanogenic substrates (Figure 3.9, 3.15). Additionally, another Stickland electron acceptor identified in our microcosm, sarcosine, was removed concomitant with the production of acetate, signifying that methylamine fermentation may contribute to acetate buildup in shales (Figure 3.11). Using our laboratory-based proteomics findings as a guide,

and corroborating to our field metabolite data, we conclude that glycine betaine fermentation is likely mediated with threonine, leucine, and glycine as possible electron donors in the field. Lysine was not detected in produced fluids, which may suggest rapid consumption in the field. Similarly, hydrogen may also represent an important electron donor that cannot be accurately measured in the field. Alternatively, we must consider the absence of measured Stickland donors in the field may signify that electron donors could be an important constraint on microbial methane production *in situ*. Collectively, our field metabolite and metagenome data signify the ubiquity of the Stickland reaction across shale well microbial communities.

Across the field produced fluid samples, microbial communities converge at late time points (>40 days after HF), despite initial differences in inoculum, well operator, or location (Figure 3.15). Thus, we next examined if the relative abundance of these produced fluid microbial communities were predictive of metabolites in the shale produced fluids. Partial least squares (PLS) regressions demonstrated that the produced fluid microbial community composition predicted the concentration of 7 metabolites in field derived fluids. These predicted metabolites included acetate, glycine, TMA, DMA chloride, ethanolamine, and glycine betaine (Figure 3.16), many of which were integral metabolites identified in our laboratory microcosms (Figure 3.11). However, ethanolamine was not included in Figure 3.14 or these remaining analyses because the correlations supporting this prediction in the field data may be spurious (Figure 3.16, Appendix B).

To better resolve the microbial strains associated with shale chemistries, we ranked the organisms' contribution to metabolite prediction using a Value Importance Projection (VIP) score to define significance (>2). A single *Halanaerobium* strain contributed to the predictions of all seven metabolites, with the top five highest VIP scores linking one strain to predictions of

chloride, acetate, glycine, TMA, and glycine betaine. This finding is consistent with our laboratory reactor inferences suggesting in saline fluids (chloride), *Halanaerobium* uses glycine to reduce glycine betaine, generating TMA and acetate. Additionally, the other three *Halanaerobium* were each predictive of different metabolite profiles, suggesting niche partitioning at the strain level may occur in this ecosystem.

Other genera identified in our laboratory microcosms also had predictive value at the field scale. For instance, for *Geotoga* the highest predictive score was for glycine concentrations, consistent with our laboratory proteomic evidence for glycine production via the glycine cleavage system. *Methanohalophilus*, which are only detected in low abundance in persisting shale microbial communities, had a strain that was predictive of glycine betaine concentrations. This is supported by our laboratory proteomics data showing the osmoprotection by these methanogens may represent a microbial source for this keystone metabolite in shales. Alternatively, this relationship to glycine betaine could be explained by the dependency of *Methanohalophilus* on glycine betaine fermentation for the synthesis of methylamine substrates. Collectively, this regression-based modeling of the field collected chemical and biological data revealed a near perfect congruence between metabolisms active in our laboratory microcosm and field scale biogeochemistry across geographically and geologically distinct fractured shales.

3.3 Conclusion

This study demonstrates how cultivation-based investigations, coupled to high-resolution meta-omics in both the laboratory and field, can help establish paradigms for microorganisms influencing terrestrial microbial ecology and biogeochemistry. Laboratory microcosms minimized many of the physical, chemical, and biological confounding factors that prevent elucidation of metabolic interactions in the field. Results from these reactors enabled us to tease

apart the complex intertwined metabolisms and trade-offs that underpin even a 'simple' microbial community (Figure 3.11). Using regression-based modeling, we show that the relative abundance of the few bacterial taxa identified in our microcosms can predict a significant fraction of the carbon and nitrogen variability in hydraulically fractured shales. The Stickland reactions identified in this study are critical to microbial persistence, providing gene targets for other protein rich environments including the human gut (38) and soils (39), where the importance of this amino acid metabolism is largely unrealized. Since our laboratory results retain their applicability at the field scale, they provide a conceptual framework to better understand or even manipulate desired biogeochemical processes in the deep terrestrial biosphere.

3.4 Materials and Methods

3.4.1 Experimental design and sample collection

Hydraulic fracturing input fluids and shale-produced fluids were collected from well heads and gas–fluid separators. These fluids were collected from 5 wells in the Utica and Marcellus shales, in Ohio (n=2), West Virginia (n=2), and Pennsylvania (n=1). Our earlier study temporally characterized geochemical and microbiological signatures of produced fluids collected from Marcellus well 1 (15). This study contributes geochemical and metagenomic data from 4 additional wells in the Utica and Marcellus shales (Appendix B-C).

In this study, a single sample from the Utica well 2 time series was used to build microcosms to assess microbial interactions among shale microorganisms. The single produced fluid sample was collected from a gas–fluid separator in October 2014 (day 96 post hydraulic fracturing) from an oil-gas well in Ohio, USA. The microcosm experiment consisted of three treatments: 1) 5mM glycine betaine and produced fluid, 2) no glycine betaine and produced

fluid, and 3) 5mM glycine betaine and no produced fluid. Each treatment was done in triplicate and consisted of 10% anoxic, produced fluid (day 96) and 90% sterile modified DSMZ 479 media dispensed in Balch tubes sealed with butyl rubber stoppers and aluminum crimps under an atmosphere of N_2/CO_2 (80:20, vol/vol). Before mixing with produced fluids, the modified DSMZ medium (per liter) included 87 g sodium chloride, 1.5 g potassium chloride, 6.0 g magnesium chloride, 0.4 g calcium chloride, 1.0 g ammonium chloride, 2.0 g yeast extract, 2.0 g trypticase peptone, 0.2 g coenzyme M, 0.2 g sodium sulfide, 4.0 g sodium bicarbonate and brought to a pH of 7.2 using 1 mM NaOH. This undefined medium was selected for two reasons (i) to facilitate sufficient biomass production necessary for proteomics measurements and (ii) to try to capture the undefined nature of many of the compounds used in the fracturing process (https://fracfocus.org/chemical-use). Growth curves were done in triplicate for each treatment, using optical density measurements at 600nm as an analog for microbial growth (Appendix B-C). Microcosm methane production was quantified at every microcosm time point that growth was measured using a Shimadzu (GC-2014) gas chromatograph equipped with a thermal conductivity detector (TCD) using helium as a carrier gas at 100°C. All GC measurements are included in Appendix C. Samples for metagenomics, metabolites, and proteomics were taken at the beginning (T_0), at maximum cell density on day 2 (T_M), and upon substantial methane production on day 20 (T_F) (Appendix B-C). To reflect the natural salinity gradient established in hydraulically fractured wells, e.g. chloride ranges from 8.3 mg/L in the input to 95 g/L over the 328 days of well sampling, our microcosms were established with a salinity of approximately 94 g/L chloride.

3.4.2 Microcosm and field fluid chemistry analysis

Twenty-one fluid samples from microcosm experiments and forty samples from Utica and Marcellus produced fluids were filtered (0.2 micron) at time of collection and sent to the Pacific Northwest National Laboratory for metabolite analysis by NMR. Samples were diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 (DSS) as an internal standard. All NMR spectra were collected using a Varian Direct Drive 600 MHz NMR spectrometer equipped with a 5 mm triple resonance salt-tolerant cold probe. The 1D ^{1}H NMR spectra of all samples were processed, assigned and analyzed using Chenomx NMR Suite 8.3 with quantification based on spectral intensities relative to the internal standard. Candidate metabolites present in each of the complex mixtures were determined by matching the chemical shift, J-coupling and intensity information of experimental NMR signals against the NMR signals of standard metabolites in the Chenomx library. The 1D ^{1}H spectra were collected following standard Chenomx data collection guidelines (143), using a 1D nuclear Overhauser effect spectroscopy (NOESY) presaturation (TNNOESY) experiment with 65,536 complex points and at least 512 scans at 298 K. Additionally, 2D spectra (including ^{1}H-^{13}C heteronuclear single-quantum correlation spectroscopy (HSQC), ^{1}H-^{1}H total correlation spectroscopy (TOCSY)) were acquired on most of the fluid samples, aiding in the 1D ^{1}H assignments of acetate, ethanol, ethylene glycol, methanol and MMA. Biological triplicates had similar metabolite pools, with all data reported (Appendix B-C). Fluid samples from the no cell control were done in single and showed consistent metabolite concentrations throughout the experiment. NMR metabolite methods and analyses of Marcellus 1 and Utica 2 produced fluids were reported previously in Daly *et al*. (15)). Here, we reanalyzed the same produced fluids to search for important compounds outlined by proteomics in the two wells

presented in Daly *et al.* (e.g. Lysine) and analyzed produced fluids from 3 additional wells (Appendix B-C). Chloride concentrations from produced fluids were obtained using a Thermoscientific Dionex ICS-2100 ion chromatograph and are included in Appendix C.

3.4.3 Metagenomic sequencing and assembly

Total nucleic acids were extracted from five microcosm samples (Inoculum (T_0), glycine betaine + cells at T_M, glycine betaine + cells at T_F, No glycine betaine + cells at T_M, and No glycine betaine + cells at T_F) using the PowerSoil DNA Isolation kit (MoBio), eluted in 100 µl, and stored at −20 °C until sequencing. DNA for the microcosm inoculum (T_0) was submitted for sequencing at the Genomics Shared Resource facility at The Ohio State University. Libraries were prepared with the Nextera XT Library System in accordance with the manufacturer's instructions. Genomic DNA was sheared by sonication, and fragments were end-repaired. Sequencing adapters were ligated, and library fragments were amplified with 5 cycles of PCR before solid-phase reversible immobilization (SPRI) size selection, library quantification and validation. Libraries were sequenced on the Illumina HiSeq 2500 platform and paired-end reads of 113 cycles were collected. The other 4 metagenomes were sequenced at the Joint Genome Institute. Briefly, libraries were created and quantified using an Illumina Library creation kit (KAPA Biosystems) with solid-phase reversible 402 immobilization size selection. Libraries were then sequenced on the Illumina HiSeq 2500 sequencing platform utilizing a TruSeq Rapid paired-end 404 cluster kit. DNA was extracted and sequenced from all produced and input fluids as outlined previously (15). All raw reads from microcosms, produced fluids, and input fluids were trimmed from both the 5′ and 3′ ends with Sickle, and then each sample was assembled individually with IDBA-UD (15, 85, 118, 144) using default parameters. Metagenome statistics including amount of sequencing are noted in Appendix B.

3.4.4 Metagenome binning and annotation for proteomics database

All scaffolds ≥2.5 kb were included when binning genomes from the metagenomic assembly. Scaffolds were annotated as described previously (15). Briefly, open reading frames were predicted with MetaProdigal (120), and sequences were compared with USEARCH (121) to KEGG, UniRef90, and InterProScan (122) with single and reverse best hit (RBH) matches of >60 bases reported. We obtained near-complete, curated draft (>93% estimated completion, <1% overages) genome resolved bins using a combination of phylogenetic signal, coverage, and GC content, for a *Halanaerobium*, *Ca.* Uticabacter, *Methanohalophilus*, and *Geotoga* (Appendix B-C). As described previously (15, 144), genome completion was estimated based on the presence of core gene sets (Bacteria, 31 genes and Archaea, 104 genes), using Amphora2 (119). Contamination (gene copies >1 per bin) indicating potential misbins, along with GC and phylogeny, were used to manually remove potential contamination from the bins. Given the dominance and high strain variation in some samples, highly abundant genomes (>400X, bacterial and viral) often failed to assemble. To recover these genomes subassemblies were performed to reconstruct the dominant genomes, using 10, 5, and 1 percent of the reads (15). Given the high strain variation, we were only able to recover a single near-complete *Halanaerobium* bin from the most abundant strain using a 1% subassembly. However, we know there were at least two other strains of *Halanaerobium* in the microcosm. In order to capture the most proteomic signal, we binned *Halanaerobium* as a whole from the inoculum to create a community *Halanaerobium* bin. This allowed us to see the activity of *Halanaerobium* as a whole in the microcosm, thus here we refer to *Halanaerobium* at the genus level. All genome statistics including 16S rRNA gene presence,

completion, and length are reported in Appendix B. Fasta files of nucleotide and amino acid sequences for each genome bin are included in Appendix C.

Near-full-length ribosomal 16S rRNA gene sequences were reconstructed from unassembled Illumina reads from microcosms and input and produced fluids using EMIRGE (145). To reconstruct 16S rRNA gene sequences we followed the protocol with trimmed paired-end reads where both reads were at least 20 nucleotides used as inputs and 50 iterations. EMIRGE sequences were chimaera checked before phylogenetic gene analyses. EMIRGE abundances for the microcosm experiment are shown in Figure 3.5 (SI Appendix). Necessary scripts and analyses to perform metagenome assembly, EMIRGE, annotation, and single-copy genes can be accessed from github (https://github.com/TheWrightonLab/metagenome_analyses).

Viral genomes were identified from all subassemblies using VirSorter (146, 147) hosted on the CyVerse discovery environment (148) (Appendix C). VirSorter was ran with default parameters using the virome database, retaining viruses and prophage with category 1 and 2 status. Viral genomes were then clustered using GenomeCluster hosted on the CyVerse discovery environment with 95% average nucleotide identity over 80% of the smallest contig (147). We combined the four microbial and sixteen unique viral genome bins to build the metagenomic database for proteomic assessment.

3.4.5 Metaproteomic extraction, spectral analysis and data acquisition

Liquid culture (1.2 ml) from each microcosm sample was collected anaerobically, centrifuged for 15 minutes at 10,000xg, separated from the supernatant, and stored at -80°C until shipment to Pacific Northwest National Lab. Proteins in the pellet were precipitated and washed twice with acetone. Then the pellet was lightly dried under nitrogen. Filter Aided Sample

Preparation (FASP) kits were used for protein digestion according to the manufacturer's instructions. Resultant peptides were snap frozen in liquid N_2, digested again overnight and concentrated to ~30 µl using a SpeedVac (Labconco, Kansas City, MO, USA). Final peptide concentrations were determined using a bicinchoninic acid (BCA) assay. All mass spectrometric data were acquired using a Q-Exactive Plus (Thermo Scientific, Waltham, MA, USA) connected to an nanoACQUITY UPLC M-Class liquid chromatography system (Waters, Milford, MA, USA) via in-house 70cm column packed using Phenomenex Jupiter 3µm C18 particles (Torrence, CA, USA) and in-house built electrospray apparatus. MS/MS spectra were compared with the predicted protein collections using the search tool MSGF+ (149). Contaminant proteins typically observed in proteomics experiments were also included in the protein collections searched. The searches were performed using ±20 p.p.m. parent mass tolerance, parent signal isotope correction, partially tryptic enzymatic cleavage rules and variable oxidation of Methionine. In addition, a decoy sequence approach (150) was employed to assess false discovery rates. Data were collated using an in-house program, imported into a SQL server database, filtered to ~1% FDR (peptide to spectrum level) and combined at the protein level to provide unique peptide count (per protein) and observation count (that is, spectral count) data. Spectral count data for each identified protein was normalized using Normalized Spectral Abundance Frequency (NSAF) calculations, accounting for protein length and proteins per sample (Appendix C). Note, metaproteomics were not done on produced fluid samples from the field.

3.4.6 Microcosm metabolic, phylogenetic and statistical analyses

Proteins for osmoprotection (Figure 3.6, Appendix B), the Stickland reaction, and other metabolisms discussed were mined from the amino acid annotation files of binned genomes

using BLASTp with a bit score cut off of 60 (a technical homolog) and cross-checked in metaproteomics data. For each metabolism discussed, scaffold and gene location for genes of interest are included (Appendix C). If >75% percent of proteins required for a multi-subunit enzyme were detected in the proteomics, we gave it the status of detected in the proteome. Significance of activity reported was based on linear discriminant analysis effect size (LEfSe) (139, 151). Linear discriminant effect size (LEfSe) analysis was performed between timepoints (e.g. T_M to T_F in glycine betaine) and treatments (e.g. T_M of glycine betaine to T_M of no glycine betaine) to find features (proteins) differentially active. LEfSe combines the standard tests for statistical significance (Kruskal-Wallis test and pairwise Wilcoxon test) with linear discriminate analysis (139). It ranks features by effect size, which puts features that explain most of the biological difference at top. LEfSe analysis was performed at the α value of 0.05 for the Kruskal-Wallis test and the threshold of 2 on the logarithmic LDA score for discriminative features. All error bars shown here are indicative of one standard deviation from the mean and all significance statements refer to a p-value of less than 0.05.

Phylogenetic analyses were performed for genome bins and metagenomes using ribosomal S3 protein amino acid sequences (genomes and metagenomes) and 16S rRNA genes (genomes only). 16S rRNA genes recovered from microcosm genomes and their nearest neighbors (SILVA database, (152)) were aligned using MUSCLE version 3.8.31. The resulting alignment was manually curated and a phylogenetic tree was constructed with RAxML 7.2.9 (GTR Gamma nucleotide model, 100 bootstrap replicates). For the S3 protein tree, amino-acid sequences were pulled from the microcosm bins and augmented with sequences mined from NCBI and JGI-IMG databases. Sequences were aligned using MUSCLE version 3.8.31 and run through ProtPipeliner, a python script developed in-house

for generation of phylogenetic trees (https://github.com/TheWrightonLab). A maximum likelihood phylogeny for the alignment of S3 ribosomal proteins and 16S rRNA genes was conducted using RAxML version 8.3.1 under the LG+α+γ model of evolution with 100 bootstrap replicates. All phylogenetic trees were visualized in iTOL (Figures 3.3-3.4.

3.4.7 Phylogenetic and statistical analysis of field data

Ribosomal S3 proteins were used to track strain resolved abundance patterns across the hydraulic fracturing dataset (Appendix B-C). First all annotated ribosomal S3 proteins from 41 input and produced fluid metagenomes were pulled to build an S3 database. Then using Bowtie 2 (123), metagenomic reads were competitively mapped by sample to the S3 database using zero mismatches. Strain resolved relative abundance was obtained by quantifying the percent of total reads that mapped divided by the length of the sequence and then normalizing to within each sample (https://github.com/TheWrightonLab). Strains included in this analysis had to have 95% of the S3 sequence covered with mapped reads. Ribosomal protein tree with all amino acid sequences used in this analysis was obtained using methods outlined above and is shown in Appendix C.

In order to predict fluid metabolites from the microcosm microbial community, we used sparse Partial Least Squares (sPLS) (153, 154) as implemented in the R package mixOmics (155). In other words, this approach enabled us to model a relationship between microbial abundance and fluid chemistry traits. In addition, the predictors were ranked according to their Value Importance in Projection (VIP) (156). The VIP measure of a predictor estimates its contribution in the PLS regression. The predictors having high VIP values are assumed important for the PLS prediction of the response variable. The VIP values of the prokaryotic functional subnetworks are provided in Appendix C.

3.4.8 Viral Analyses

We used two methods to link viral contigs to microbial hosts. First, as described previously, CRISPR arrays were identified in each genome bin by using the CRISPR recognition tool plugin in Geneious R8 (157). To link microbial hosts and viruses, we used BLASTn to identify viral contigs with matching spacer sequences. All matches were manually confirmed as perfect matches by aligning sequences in Geneious R8. Second, we used the d_2^S hexamer frequency dissimilarity measure (158) between viral contigs and host genomes to predict viral-host associations. Analyses were run with 5 microcosm genomes and 16 viral population representatives. In all cases, the d_2^S dissimilarity measure predictions were congruent with CRISPR spacer array linkages.

In Figure 3.7, expressed viral proteins are divided into 7 categories: DNA/Replication, Lysogeny, Structure, Lysis, hypothetical, transposase, and other. DNA/replication category referred to amino acid sequences associated with DNA metabolism such as DNA methyltransferases and helicases. Lysogeny refers to the viral lysogenic cycle and was made up of recombinases and resolvases. The structural category included tail sheath proteins, terminases, and phage tail tape measures. The transposase category was only made up of transposase associated amino acid sequences, while hypothetical referred to proteins of unknown function or hypothetical distinction.

Table 3.1 Summary of Stickland half reactions.

Compound	Relevant gene	Half Reaction	Donor/Acceptor
GB	GB reductase *(grdHI)*	GB ⇒ Acetyl-P + TMA	Acceptor
Sarcosine	Sarcosine reductase (*grdFG)*	Sarcosine ⇒ Acetyl-P + Monomethylamine	Acceptor
Glycine	Glycine reductase (*grdBE)*	Glycine ⇒ Acetyl-P + Ammonium	Acceptor
Lysine	3,5 diaminohexanoate dehydrogenase (*kdd*)	Lysine + NAD^+ ⇒ ⇒ 3,5 diaminohexanoate + NADH	Donor
Threonine	Threonine dehydrogenase (*tdh*)	Threonine + NAD^+ ⇒ L-2-amino-3-oxobutanoate + NADH	Donor
Glycine	Glycine dehydrogenase (*gvcD*)	Glycine + NAD^+⇒ Ammonium + CO_2 + NADH	Donor

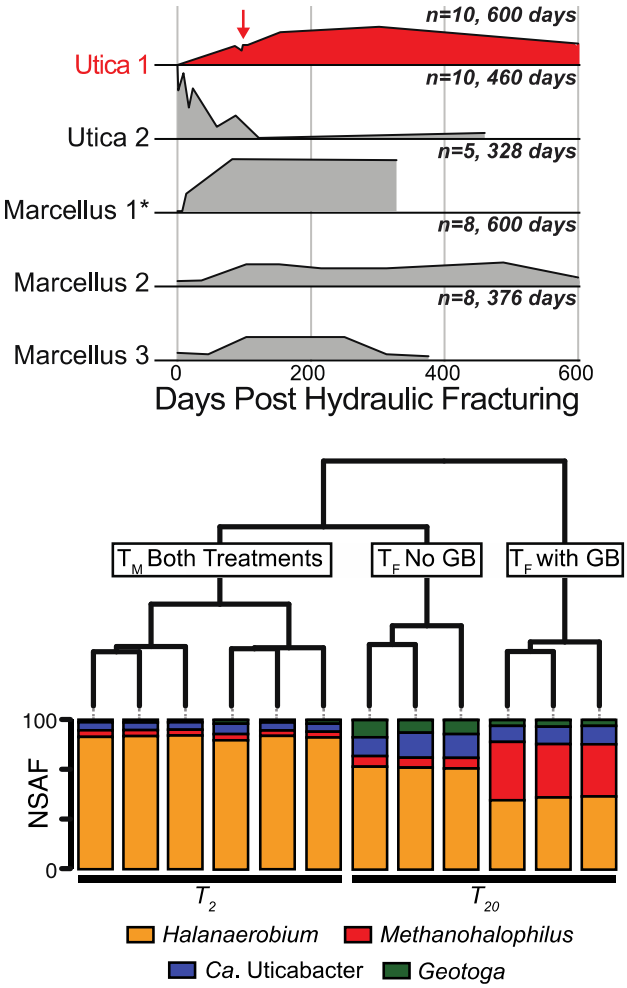

Figure 3.1 Glycine betaine concentrations across five wells.

Area plots show glycine betaine concentration trends through time across five HF wells in two different shale formations, with the number of samples denoted. Inocula for microcosm experiments were obtained from the well shown in red at the time point indicated by the red arrow. Data is shown from a previous study (Daly, *et al.*) and is indicated by the (*).Hierarchal clustering of microcosm experiment metaproteomes is shown for detected proteins from 50 hour (T_2) and 425 hour (T_{20}) time points. Stacked bars below represent each metaproteome with relative abundance of proteins from each organism indicated by color within each sample. Time point and microcosm treatment are indicated in black and grey below, respectively.

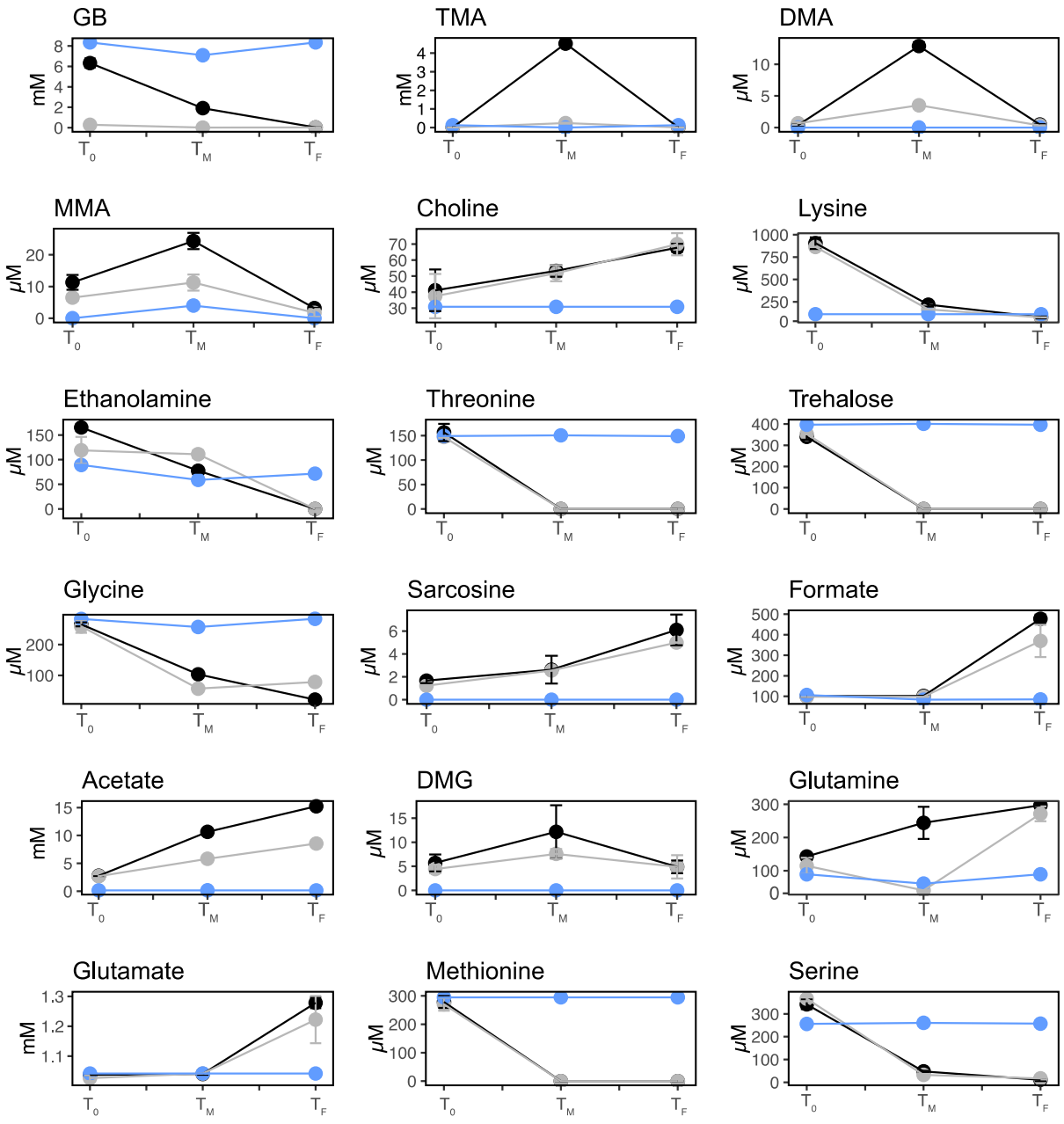

Figure 3.2 Microcosm metabolite concentrations over time.

Graphs show all metabolites detected in microcosms by NMR, with all treatments shown (with glycine betaine= Black, no glycine betaine= grey, Media Control= blue). Points indicate triplicate average and error bars show one standard deviation from the mean.

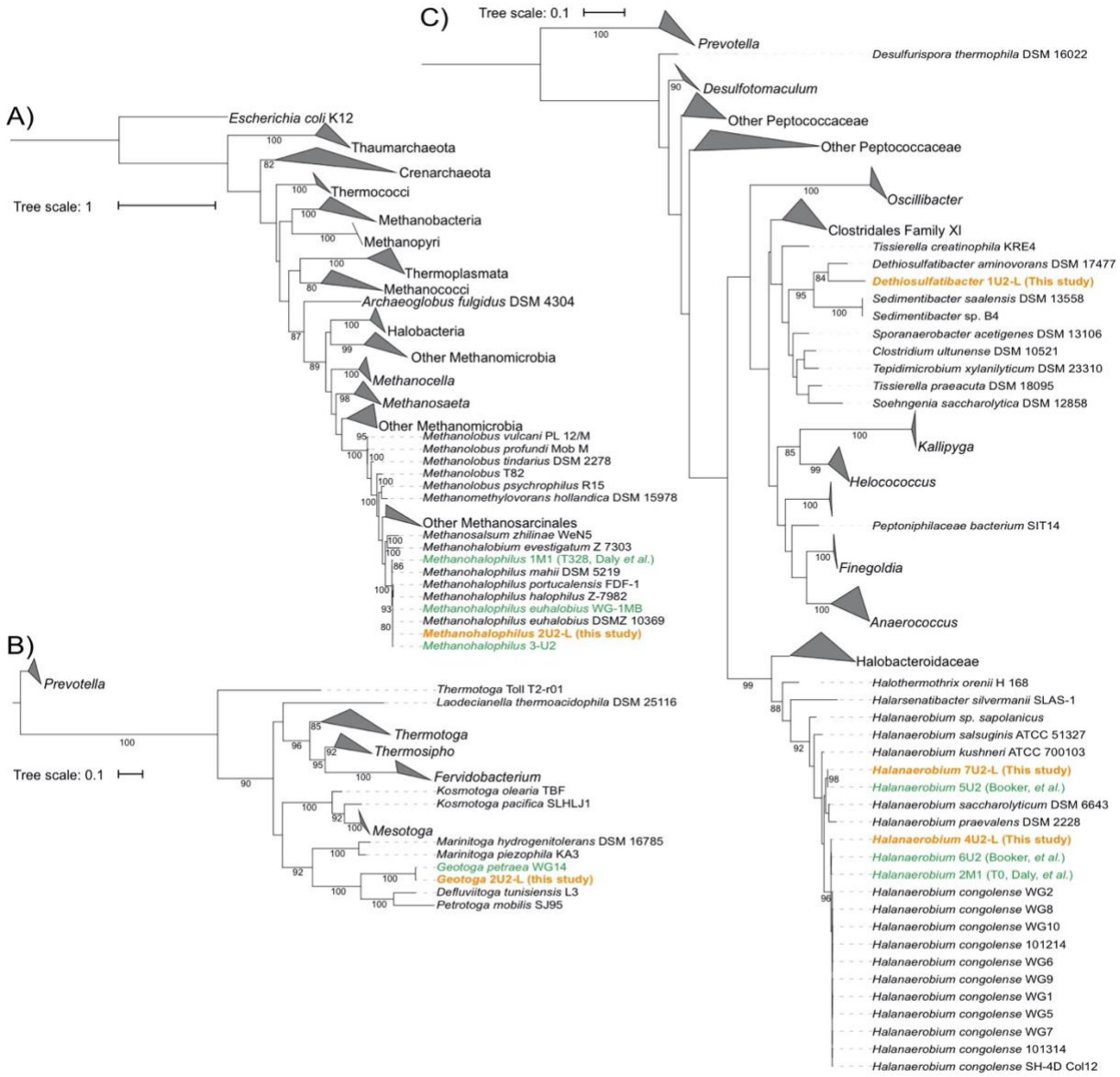

Figure 3.3 Ribosomal S3 protein trees.

Maximum likelihood S3 ribosomal protein trees of archaea (A) and Thermotogaceae (B), and Firmicutes (C), showing the taxonomic assignment of genomes from the microcosm experiment. S3 amino acid sequences from bins in this study are shown in orange, while sequences from (4) are shown in green.

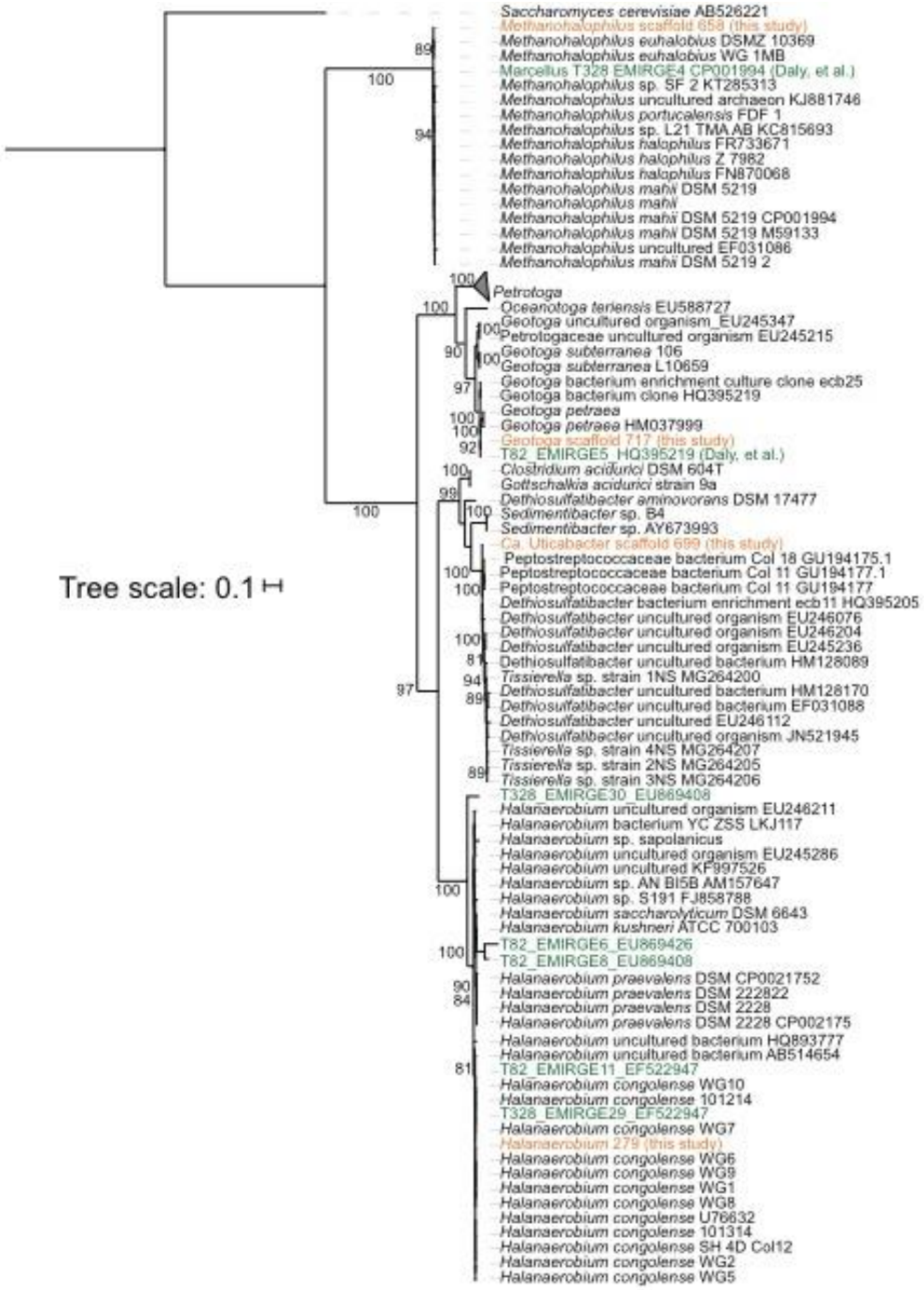

Figure 3.4 16S rRNA gene tree.

Maximum likelihood 16S rRNA tree, showing the taxonomic assignment of genomes from the microcosm experiment. 16S rRNA from bins in this study are shown in orange, while sequences from (4) are shown in green.

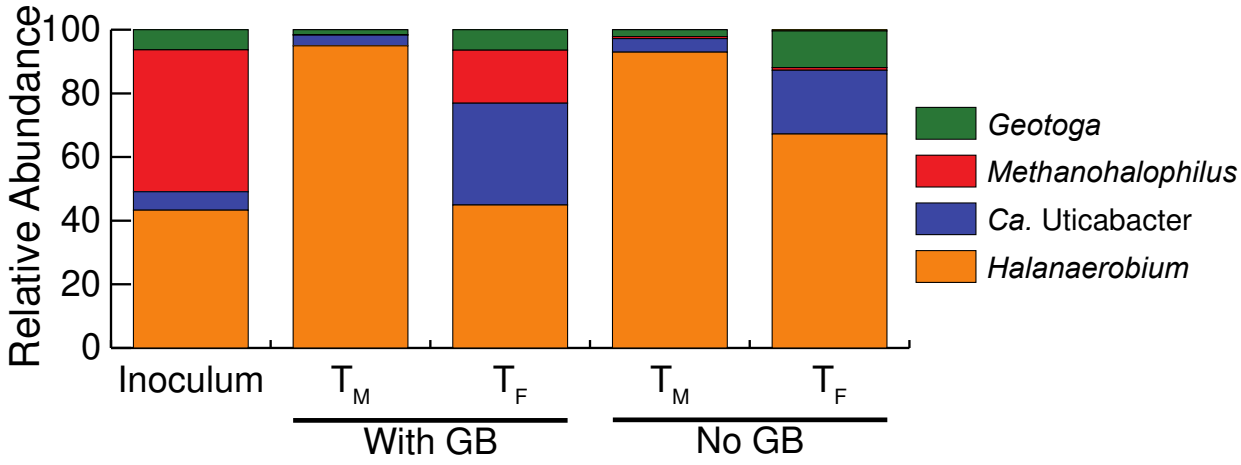

Figure 3.5 EMIRGE relative abundance in microcosms.

Relative abundance by EMIRGE of all time points in the glycine betaine and non-glycine betaine microcosm experiments. Stacked bars are colored by organism within each metagenome. Organisms with >0.05% relative abundance are shown.

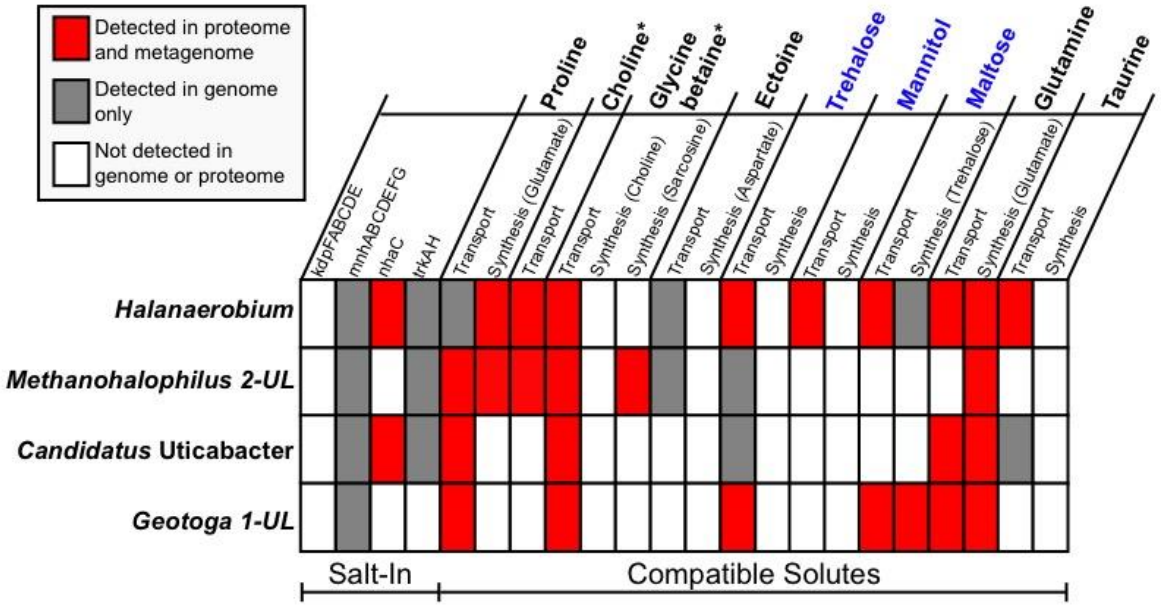

Figure 3.6 Osmoprotection strategies utilized by microcosm microbial community.

Heatmap denotes active and potential osmoprotection strategies utilized by microcosm microbial community. Both salt-in and compatible solute strategies are considered. Compatible solute compounds found in the produced fluid Utica well time series are denoted with an asterisk (*), while sugar compatible solutes are shown in blue text. For multisubunit enzymes, >75% percent of proteins were required for detected in the proteome status.

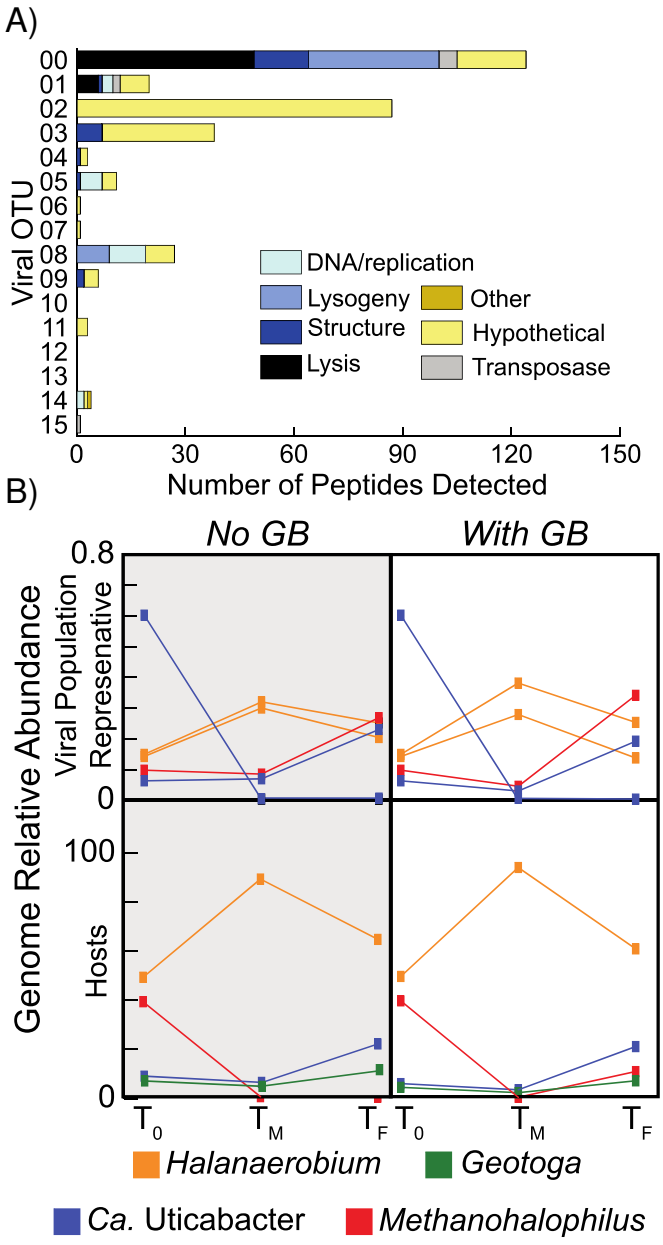

Figure 3.7 Viral peptide abundance in microcosms.

(A) Stacked bar chart denotes detected unique viral peptides per representative genome broken into 7 different categories (see methods). (B) Genome relative abundance of microbial hosts (bottom) and viral population representatives (top) are shown across time and within treatments (No glycine betaine and With glycine betaine shown on the left and right, respectively). Only viral populations with >0.1% relative abundance in at least one timepoint in glycine betaine microcosm are shown. Viral OTUs represented in B include 00 (*Methanohalophilus*, red), 02 (Ca. Uticabacter, blue, decreasing from T0 to TF) ,10 (*Halanaerobium*, orange, least abundant), 13 (Ca. Uticabacter, blue, increasing from T0 to TF), and 15 (*Halanaerobium*, orange, most abundant).

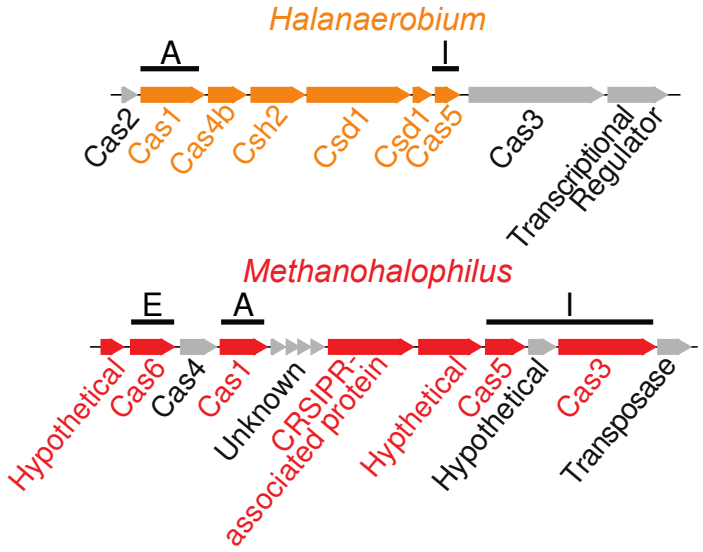

Figure 3.8 *Halanaerobium* and *Methanohalophilus* CRISPR-Cas systems.

Halanaerobium and *Methanohalophilus* CRISPR-Cas system genes are shown, with corresponding peptides detected in proteomics highlighted in orange and red, respectively. Genes for adaptive immunity are denoted by functional stage, with Adaptation (A), Expression (E), and Interference (I) stages all represented in metaproteomic data.

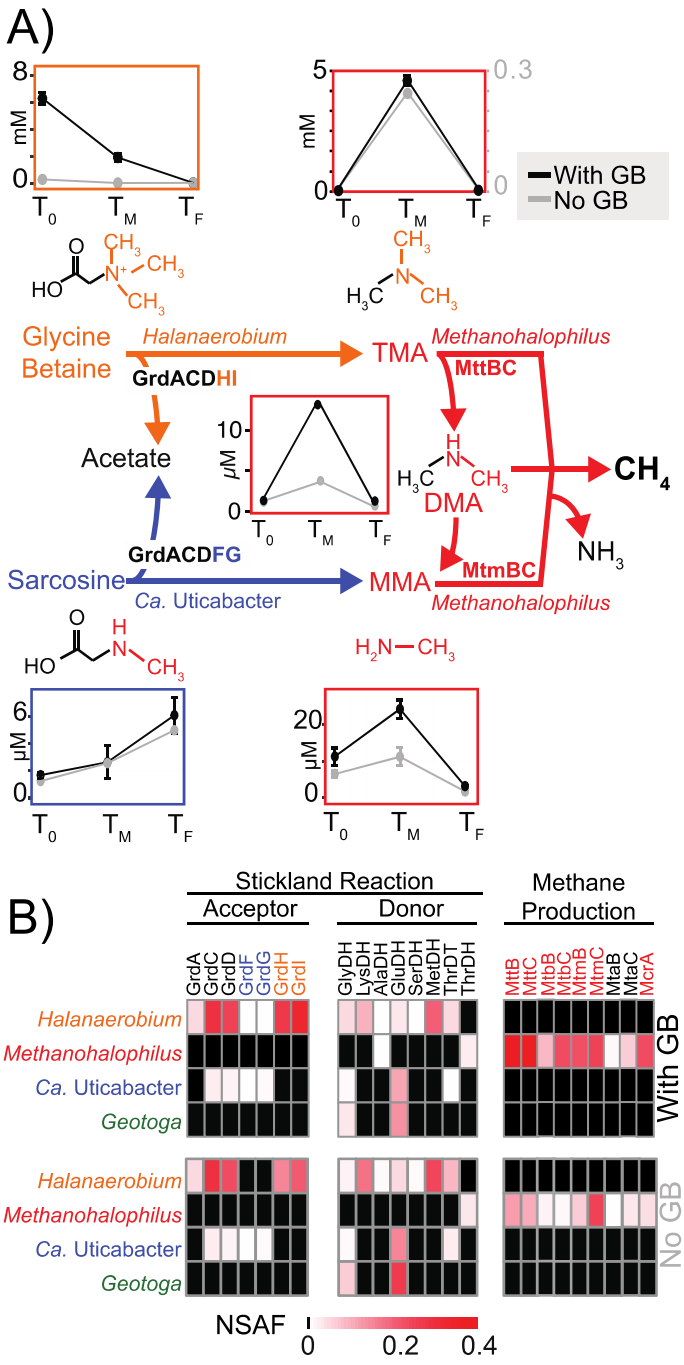

Figure 3.9 Metabolite and metaproteomic evidence for Stickland reactions in microcosms.

(A) Center colored pathway shows Stickland reactions from glycine betaine and sarcosine to TMA and methylamine (MMA) respectively, fuel methanogenesis with pathways colored by organism. Chemical structures are shown, with cleaved products colored. Corresponding line graphs shows average metabolite concentrations with standard deviation of triplicate samples through time colored by treatment (black= glycine betaine and grey= No glycine betaine). Note, TMA is reported with a dual y-axis and all dynamics of methanogenesis substrates (TMA, DMA, and MMA) are shown in red boxes and acetate concentrations overtime can be found in Figure

3.9. *Geotoga* is not represented here because it does not have the potential to carry out a Stickland reaction. (B) Heat maps display NSAF values for proteins detected by metaproteomics in glycine betaine amended (top) and No glycine betaine (bottom) microcosms at the TF timepoint.

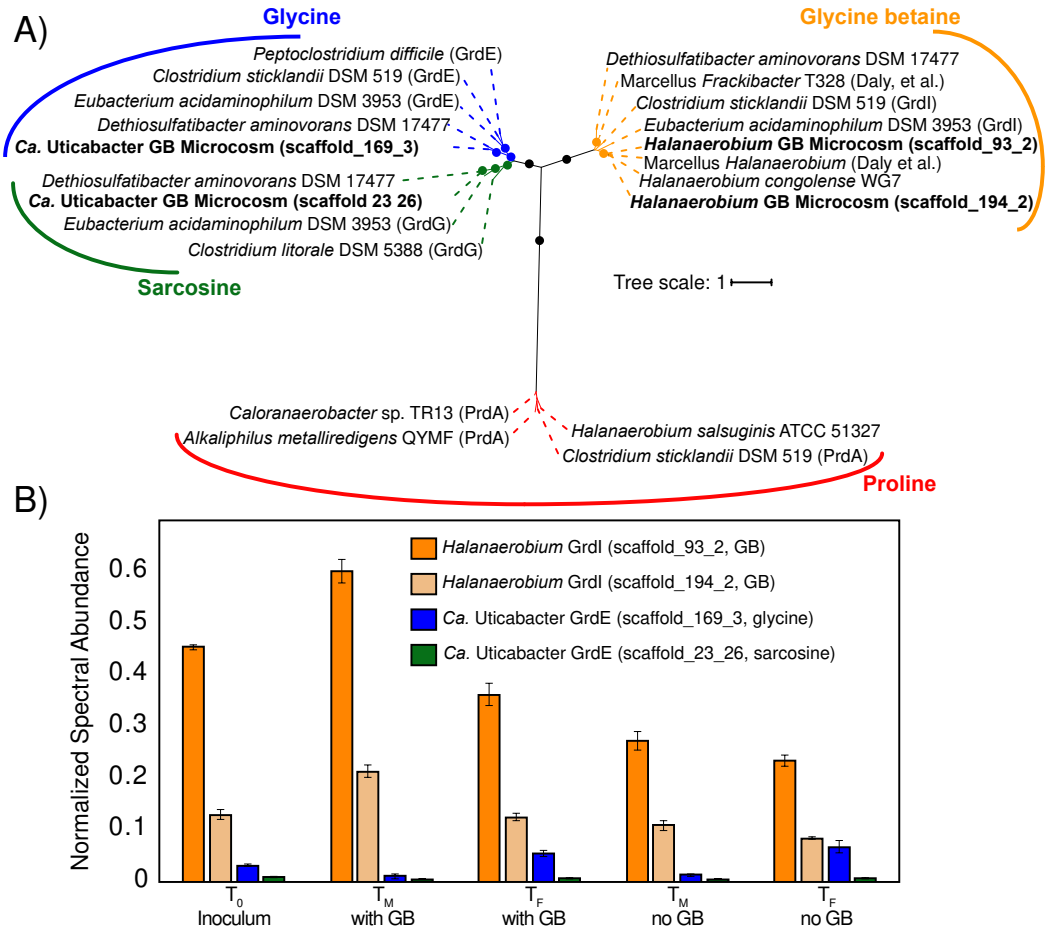

Figure 3.10 Activity of reductase systems for glycine, glycine betaine, and sarcosine in microcosms.

(A) Phylogenetic analyses of GrdE (blue, glycine), GrdG (green, sarcosine), GrdI (orange, glycine betaine), and PrdA (red, proline) proteins from microcosm experiments showed that proteins clustered by substrate specificity. *Halanaerobium* had two active copies of glycine betaine reductase and *Candidatus* Uticabacter had an active glycine and sarcosine reductase, as these formed monophyletic clades with known glycine betaine, with known reducers of the respective methylamine substrates, *Eubacterium acidominophilum* and *Clostridium sticklandii*. Sequences from this study are in bold and include the genome and scaffold number followed by the relevant gene number(s). Bootstraps >90 are shown with closed circles at nodes. (B) Proteomic expression of reductases in (A) is shown by time point (x-axis) with color denoting reductase mechanism. Bars represent average activity of biological triplicates with standard deviation shown (error bars).

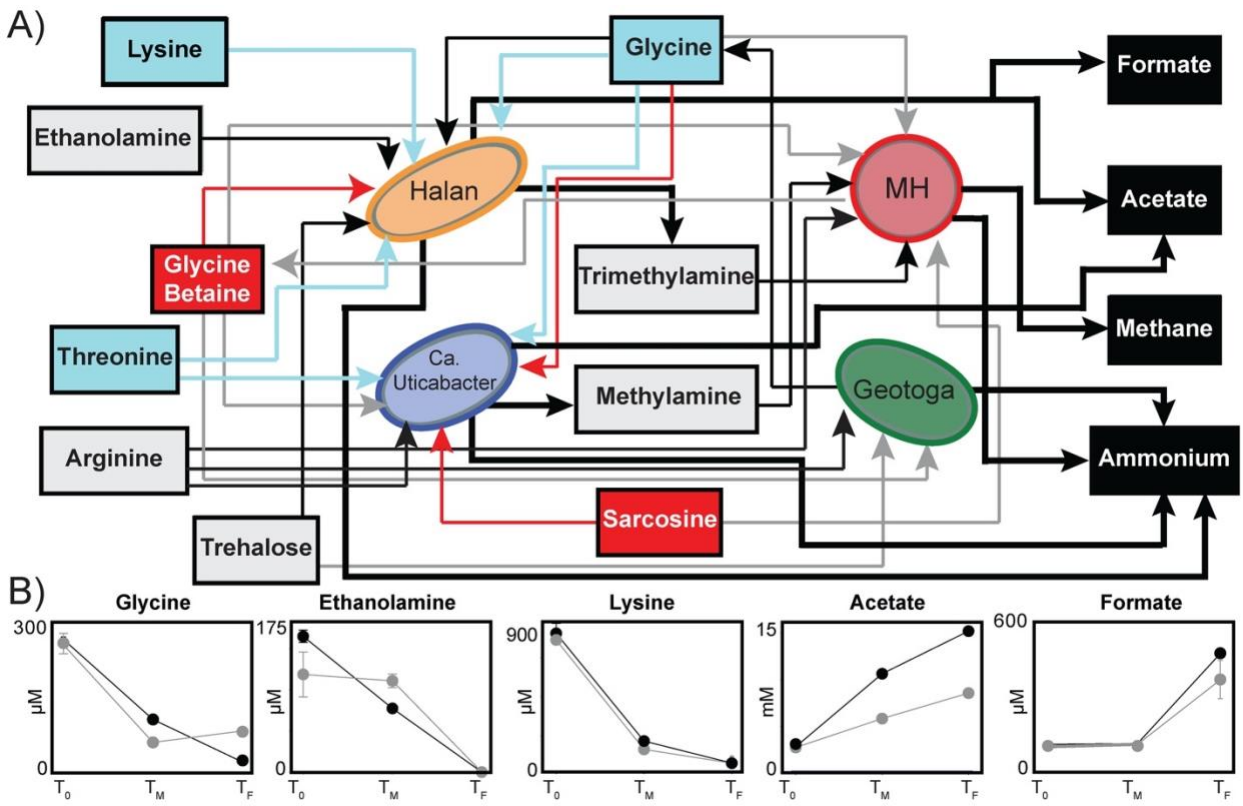

Figure 3.11 Metabolic network of interactions revealed by metaproteomics and metabolite analyses.

(A) Network of *Halanaerobium* (orange), *Methanohalophilus* (red), *Ca.* Uticabacter (blue), and *Geotoga* (green) shows the interconnected metabolisms of shale organisms. Arrows pointing toward and away from microbes show utilization and production, respectively. Arrow line color denotes substrate utilization: red (oxidant in the Stickland reaction), blue (reductant in the Stickland reaction), grey (osmoprotectant). Bold black lines indicate the production of substrates and terminal end products are noted in black boxes. (B) Line graph shows average with standard deviation of triplicate metabolite concentrations through time colored by treatment (black= glycine betaine and grey= No glycine betaine). Abiotic control metabolite concentrations did not change significantly over time but showed glycine was added from media not produced fluids (Figure 3.2, Appendix C).

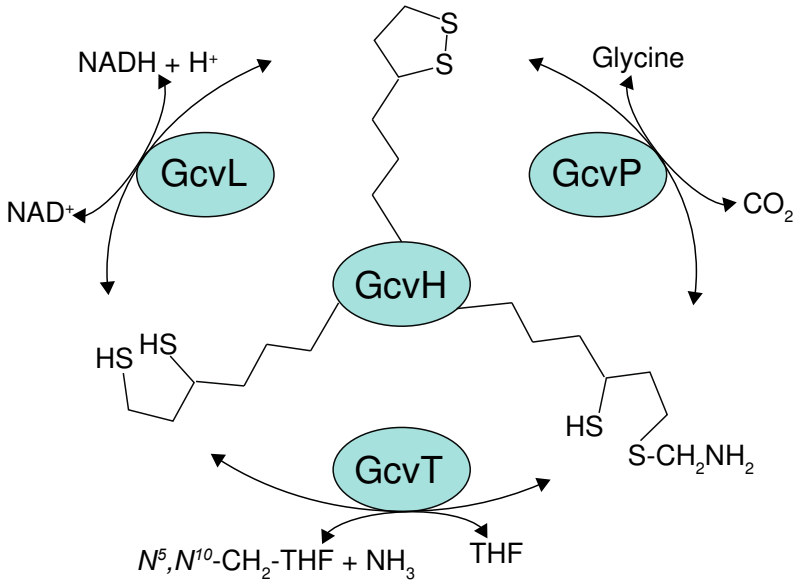

Figure 3.12 Glycine cleavage system.

Schematic of the glycine cleavage system (Gvc) is shown, with all of the enzymes denoted by blue ovals detected in metaproteomics.

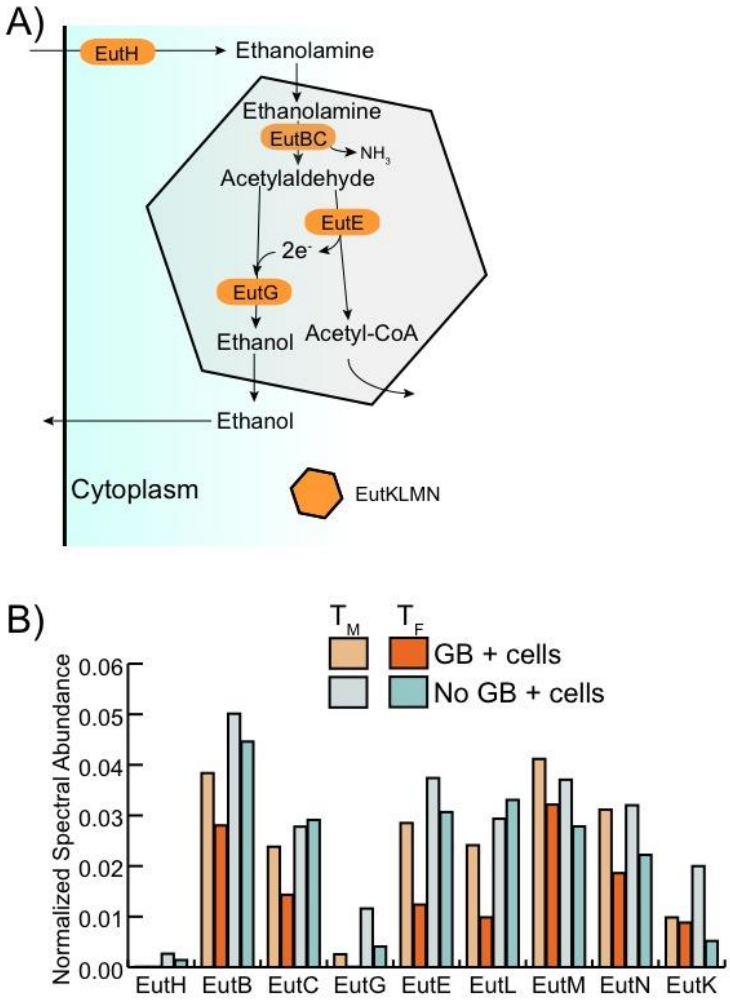

Figure 3.13 Ethanolamine utilization in *Halanaerobium*.

(A) Pathway of ethanolamine utilization by *Halanaerobium.* All proteins shown were detected in the metaproteomics. (B) Relative quantification of ethanolamine utilization proteins. Each bar represents the average NSAF value for each protein (in triplicate) within each time point by treatment.

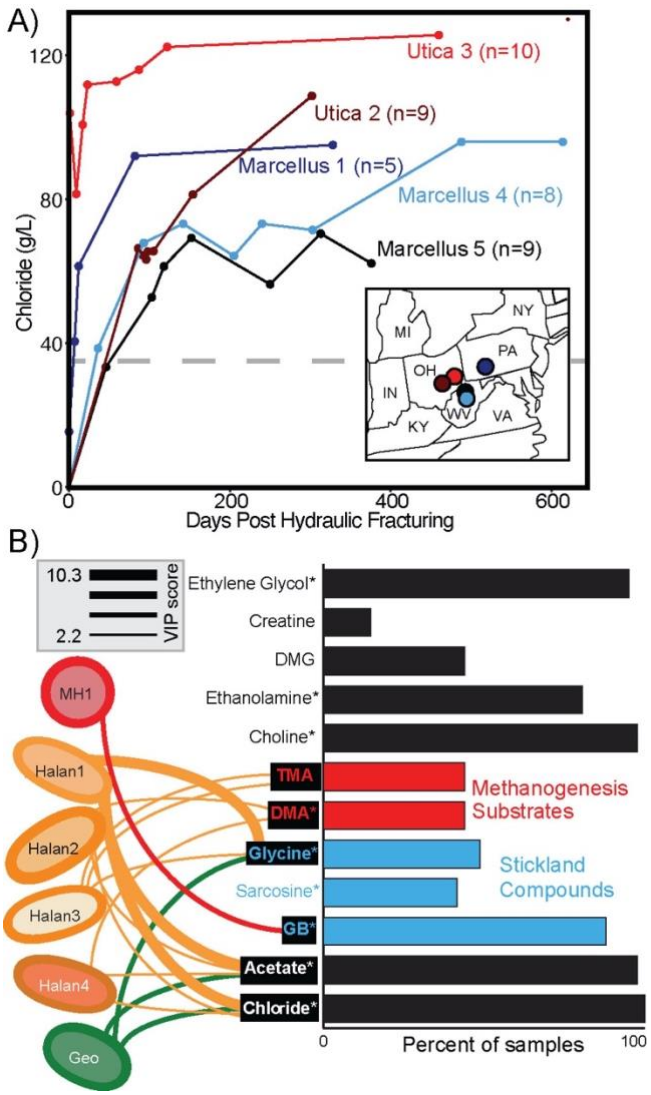

Figure 3.14 Predictions of field metabolite data with microbial abundance.

(A) Chloride concentrations for field samples with paired metabolites and metagenomes are shown (n=41), with color denoting well. Above the dashed line indicates hypersaline conditions. Circles on the inserted map show each well's geographic location. (B) Bar graph shows the prevalence of key metabolites uncovered by laboratory experiments across 41 input and produced fluid samples from 5 wells. Substrates are colored by metabolism (red= methanogenesis substrates, blue= Stickland reaction substrates). Asterisks signify metabolites detected in at least one of the input fluids described here. Concentration of field metabolites that could be significantly predicted (sPLS regression, R^2>0.3). by the field relative abundance of microorganisms are denoted with black boxes. Taxa from microcosm experiments that were significant variables (VIP values >2) in metabolite prediction are shown by connections between metabolites and *Halanaerobium* (Halan), *Methanohalophilus* (MH), and *Geotoga (*Geo), with the thickness of the line denoting variable importance. Top 3 predictions are shown for each strain, with *Halanaerobium* strains numbered 1-4.

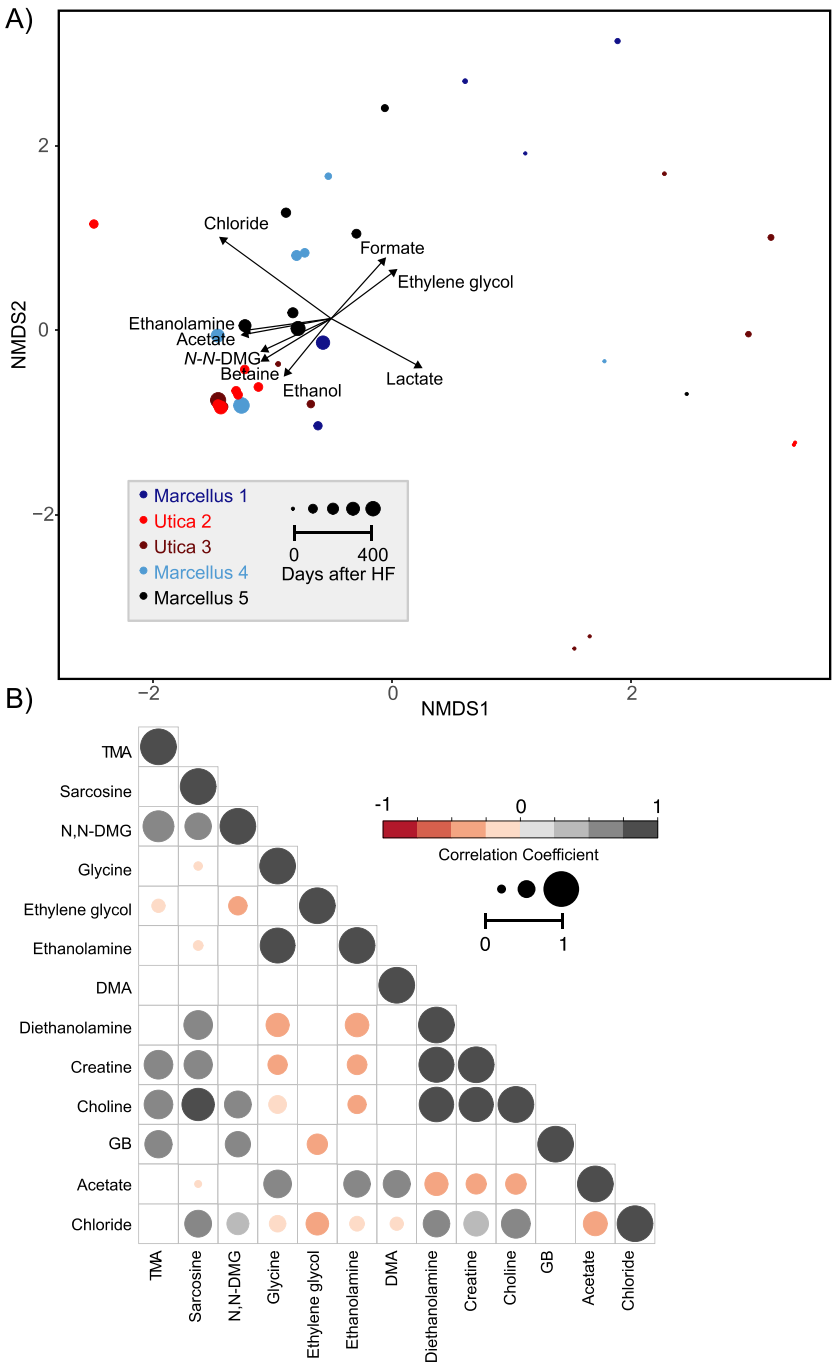

Figure 3.15 Field scale metabolite correlations.

(A) Nonmetric multidimensional scaling (NMDS) of microbial community abundance overlaid with geochemistry. All vectors show significant associations between microbial communities and paired sample chemistry (envfit, p-value<0.05). Samples are colored by well and bubble size denotes time after HF. (B) Bubble plot shows significant correlations between metabolites analyzed by NMR in 26 produced fluid samples collected from five HF wells, inputs were excluded from the analysis for clarity. Bubble color and size denotes correlation coefficients using colored scale bar below.

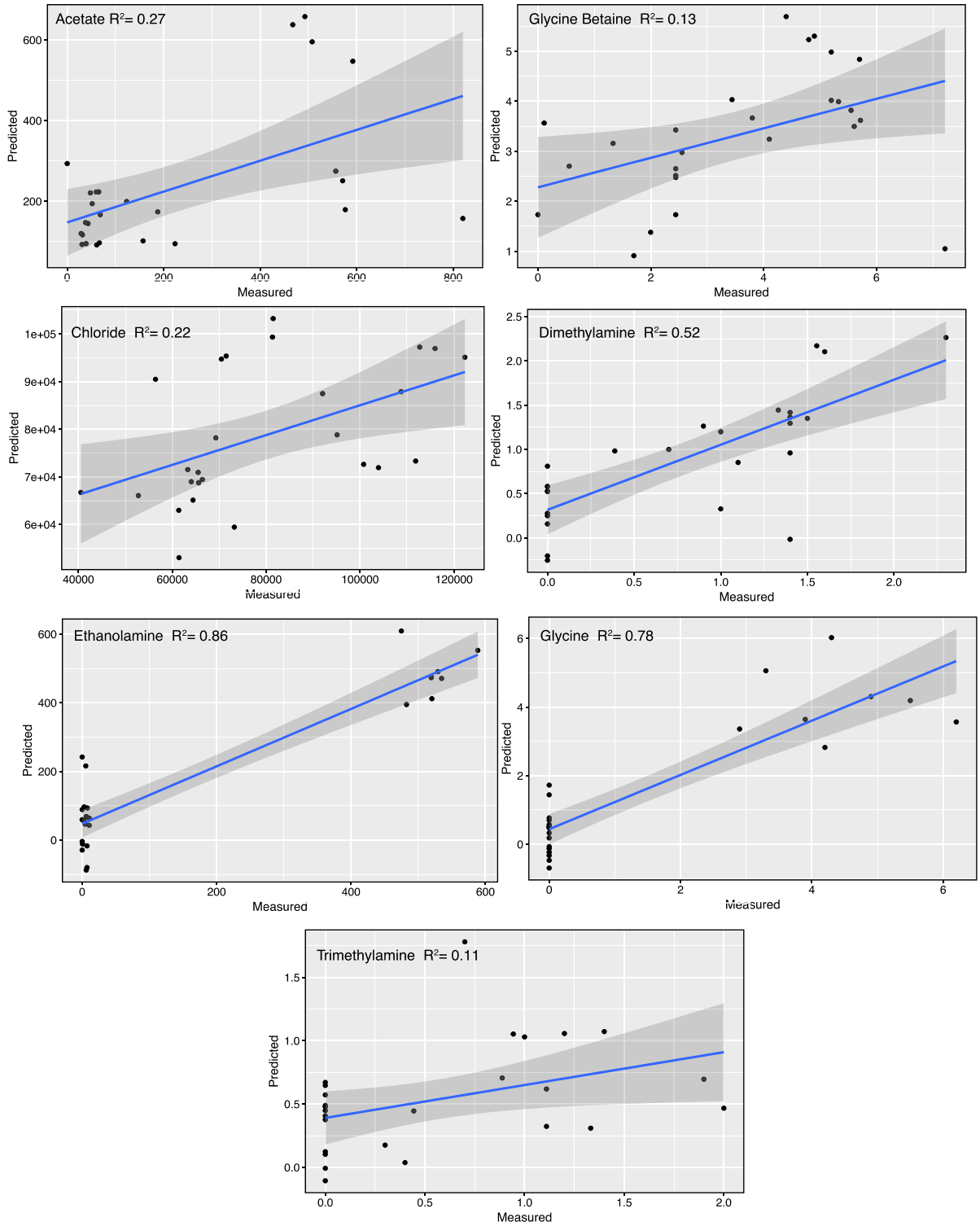

Figure 3.16 Correlations of measured versus sPLS predicted metabolites at field scale.

Produced fluid microbial communities predict acetate, chloride, dimethylglycine, glycine, trimethylamine, and glycine betaine using sPLS. Correlations of measured versus predicted for each metabolite is shown (all p-values < 0.05).

Chapter 4: Microbial methylated amine metabolism in the gut is predictive of cardiovascular disease in humans

4.1 Introduction

Mounting evidence implicates the gut microbiome as a critical component of human health and disease, with a well-defined illustration of this relationship being the link between gut microorganisms and atherosclerotic cardiovascular disease (ACVD) (5–7). In particular, methylated amines (MAs), nitrogen-containing compounds bearing one or more methyl groups, are key metabolites in this disease-relevant metabolic network. Specifically, extensive evidence from mouse and human studies showed that trimethylamine-N-oxide (TMAO) has a strong clinical prognostic value for ACVD and is derived from a microorganismal pathway fueled by the human diet (5, 7, 47, 159, 160). This research showed that microorganisms convert diet-derived quaternary amines, such as choline, glycine betaine, carnitine, and butyrobetaine to the proatherogenic metabolite trimethylamine (TMA) (Figure 4.1A, orange), a metabolite exclusively produced by the gut microbiota (5, 7, 47, 159). From the gut, TMA enters the bloodstream and host liver enzymes oxidize it to TMAO, a compound that triggers macrophage-mediated lipid deposition and ultimately ACVD (160, 161). More recently, other biochemical evidence revealed alternative pathways for quaternary amine degradation that do not produce TMA, in which quaternary amines are instead demethylated to nonatherogenic by-products such as dimethylglycine and norcarnitine (Figure 4.1A, green) (25, 28, 32). By leveraging biochemical knowledge uncovered within the last two decades, it has become possible to piece together a holistic view of microbial MA metabolism in the gut.

Using a combination of methods, proatherogenic (TMA-producing) and nonatherogenic (not TMA-producing) enzymes and their respective genes were recently discovered, and a handful of model microorganisms characterized (22, 25, 27, 28, 32, 34) (Figure 4.1B). In general, proatherogenic enzymes cleave quaternary amines to release TMA, whereas nonatherogenic enzymes demethylate quaternary amines and trimethylamine. TMA-producing, proatherogenic genes include *cutC* (choline-TMA-lyase) discovered in *Desulfovibrio desulfuricans*, *grdI* (glycine betaine reductase) in *Eubacterium acidaminophilum, cntA* (carnitine monooxygenase) in *Acinetobacter baumannii*, and *yeaW* (butyrobetaine monooxygenase) in *Escherichia coli* (22, 27, 33, 34) (Figure 4.1B). Alternatively, nonatherogenic enzymes for the demethylation of MAs have been identified across Bacteria and Archaea (25, 28, 32, 46). These non-atherogenic genes are comprised of methyltransferases that do or do not encode for amino acid pyrrolysine, making automated annotation difficult (31). Particularly, genes belonging to the MttB superfamily of trimethylamine methyltransferases that encode for pyrrolysine demethylate trimethylamine (*mttB*), while those that do not, have been shown to demethylate quaternary amines (*mtcB* and *mtgB*) (10, 11, 28). Likewise, dimethylamine (*mtbB*) and monomethylamine (*mtmB*) methyltransferases, first discovered in methylotrophic methanogens, also encode for pyrrolysine (Figure 4.1A) (10, 11). While these detailed studies delivered key biochemical knowledge, there remains to be a study that integrates these features into a comprehensive metabolic network, applying it to large scale datasets. Limited study of MA metabolism in the gut is likely due to the difficulty in accurate annotation of these genes. Bottlenecks in annotation occur due to lack of characterization in public databases, specificity often depends on single amino acid level differences, or genes are truncated due to pyrrolysine which are mistakenly assigned as stop codons during gene calling (22, 27, 28, 31).

Cultivation independent investigations of MA metabolism have been gene-focused rather than genome focused or considered only proatherogenic TMA producing pathways (29, 30, 162–165). The best profiled methylamine gene across datasets is *cutC*, with investigations demonstrating that the potential for choline degradation to TMA is prevalent in 96% of HMP stool metagenomes (162). The impact of methylotrophic methanogens in the human gut has also been considered, finding that the presence of Methanomassiliicoccales *mttB* was associated with lower concentrations of fecal TMA (30). However, this same study also found an increased number of TMA producing pathways (*cutC*, *grdI*, and *cntA*) was not significantly associated with higher fecal TMA concentrations (30). This suggests that TMA is subject to metabolic handoffs across the gut microbiome and indicates that methylamine genes must be investigated within the context of an integrated metabolic network as well as activity measurements in order to understand human health outcomes. Beyond methylamine gene surveys, 16S rRNA gene and metagenomic studies examining the gut microbiome of patients with and without ACVD or with elevated levels of circulating TMAO have shown increased levels of *Collinsella*, *Klebsiella*, *Escherichia*, *Prevotella*, *Peptostreptococcaceae, Clostridium, and Fusibacter,* as well as increased potential for TMA-production, peptidoglycan synthesis, and degradation of fatty acids (6, 29). These studies also hint at genome-level differences within specific genera that may influence ACVD, as members of the genus *Eubacterium* have been associated with both control and ACVD patients in separate studies (7, 29, 163, 166). Despite these intellectual and methodological advances, published studies do not use genome-resolved approaches, measure microbial activity, or consider microbial interactions (e.g. cooperation and competition for substrates). As a result, a systems-level understanding of MA metabolism is lacking in the human gut.

To address this knowledge gap, we built a Gut-Associated Methylated Amine database (GAMAdb) comprised of 6,341 microbial genomes from 13 phyla that encode genes for MA metabolism. Data for GAMAdb was collected through our metagenomic reconstruction of hundreds of near-complete genomes from 54 human gut metagenomes, and also mined from 237,273 previously published metagenome-assembled and isolate genomes from the human gut (167, 168). This database was coupled to metaproteomic and metabolomic time series data from methylamine fed gut laboratory reactors that not only demonstrate the activity of quaternary amine degradation, but also indicate MA metabolism is an emergent property in gut microbial communities. Beyond laboratory studies, human gut multi-omic analyses demonstrated predictability of disease relevant host metabolites (e.g. trimethylamine) and cardiovascular disease from microbial gene abundances. These datasets and analyses provide unprecedented, comprehensive insight into the diversity and ecology of MA metabolism that ultimately unveils key predictors of ACVD in the gut microbiome.

4.2 Results and Discussion

4.2.1 Fecal microbiota membership predicts host metabolite concentrations

To characterize the gut microbiome in relation to quaternary amine degradation, we collected paired fecal and urine samples from 125 subjects at a single time point for microbial community and metabolite analysis (Figure 4.1C). After removal of subjects who failed to meet study inclusion criteria (Methods), our final cohort consisted of 113 subjects, from which we collected paired fecal and urine samples at a single time point for chemical and microbiological analyses (Figure 4.1-4.2, Appendix D). This cohort included 76 males and 37 females self-described as healthy, with an average age of 42.3 ± 11.7 years (Figure 4.1DE). Cigarette smoking was reported by 60% of males and 40% females (Figure 4.1D), a smoking status

comparable to metabolite studies performed in this region of the United States (169). Factoring height and weight together, the cohort had an average Body Mass Index (BMI) of 26.4 ± 5.7, with 29% classified as overweight (BMI>25) and 21% classified as obese (BMI>30) (Figure 4.1DE, 4.3). The dietary meat consumption and mean body mass index of this cohort is consistent with the American average (170, 171).

We investigated the relationship between host factors and the distribution of MA metabolites. From paired fecal and urine samples from each subject (n=113), we performed targeted metabolite analysis of quaternary amines (derived from the diet and microbial transformations), TMA (the key proatherogenic metabolite produced exclusively by the gut microbiota), and TMAO (the host produced metabolite that is prognostic of ACVD). Of the four quaternary amines, glycine betaine had the highest average (6.9 μM/mM ± 27.5) and maximum concentration (283.1 μM/mM), while butyrobetaine had the lowest average concentration (0.9 μM/mM ± 1.2) (Figure 4.4A, Appendix D). Across the cohort, fecal TMA ranged from below detect to 570 nmol/gram feces, while TMAO ranged from below detection to 174.3 μM/mM in the urine (Figure 4.4A, 4.5). The average concentration and variance of these metabolites did not differ by sex or other host criteria (Figure 4.6).

Our analyses failed to identify a significant relationship of a subject's MA profile and sex, BMI category, or smoking status (Figure 4.6). Interestingly, the only non-atherogenic demethylation product we measured, dimethylglycine, was positively correlated to its precursor glycine betaine, while no relationship was observed between glycine betaine and TMA. This finding suggests the gut microbial community more readily demethylates this gut quaternary amine to an innocuous metabolite, rather than reduce it to the proatherogenic TMA. Our

collective metabolite results showed that quaternary amines and their microbial degradation products are prevalent and variable across this human cohort (Figure 4.5).

To assess the integrated diet–microbial host cometabolism of MAs, we performed 16S rRNA gene sequencing on cohort fecal samples, analyzing microbial community data together with metabolite and host metadata. Consistent with our metabolite data, microbial community membership, diversity, and richness were not statistically different by methylamine profile, gender, smoking status, or BMI (Figure 4.7). Given that this metabolism may not be broadly encoded, we hypothesized that groups of organisms, rather than features of the entire community, could be correlated to MA metabolites. We built a network where nodes were microbial taxa (16S rRNA Amplicon Sequence Variants, ASV) and links represented the co-occurrence of these taxa across our cohort. This network was then clustered into modules (n=39) that were correlated to host metadata to examine significant module–methylamine relationships. Nearly a fourth (n=9) of these modules showed a significant positive correlation to quaternary amines, TMA, or TMAO concentrations, of which, three modules (6,8,9) were predictive of fecal TMA (153), host TMAO concentrations, and even a subject's meat consumption, further supporting a relationship between diet and the enrichment of specific bacterial taxa (47, 170) (Figure 4.4B-G). Analysis of the membership in these methylamine related modules was consistent with prior reports inferring positive associations between gut microbial taxa and ACVD risk or onset (Figure 4.4D, 4.8). For example, in mice and humans, members of *Prevotella,* Lachnospiraceae*,* and *Ruminococcus* were associated with circulating TMAO concentrations, while *Ruminococcus* had greater relative abundance in ACVD individuals (7, 29, 163, 166). Moreover, *Eubacterium limosum* is a model organism for demethylating quaternary amine to nonatherogenic products, with demonstrated growth on carnitine, butyrobetaine, and

glycine betaine (25, 172) . Here we expand these findings to show that specific groups of taxa can predict MA concentrations in the gut, yet despite this perceived importance, the mechanisms underpinning these relationships are not yet realized.

4.2.2 The GAMA-gene database comprehensively catalogs the myriad of methylamine enzymes encoded by the gut microbiome

To develop a strategy for unveiling the biochemical repertoire encoded by the human gut microbiome, we constructed the Gut Associated Methylated Amine (GAMA) database (Figure 4.9AB, Appendix D). Here we cataloged the microbial methylamine related genes in the feces of our cohort, including the metagenomic sequencing, assembly, and binning of DNA extracted from (i) 54 subjects and from (ii) 5 samples from methylamine-enriched laboratory reactors established with feces from our cohort. We also combined our data with (iii) 700 genomes from isolates in the Human Microbiome Project (HMP) and (iv) 237,273 gut derived metagenome-assembled genomes (MAGs) from previously published studies (167) and (168). These later MAGs were compilation studies, where MAGs were accumulated across many publications representing many different lifestyles, disease types, and diets (167, 168). This entire genome collection (n=238,530) was computationally mined for genomes containing proatherogenic or nonatherogenic genes (Methods, Figure 4.10) to create first of its kind GAMA genome-resolved and gene-resolved databases.

Despite their intriguing connections to human biology, and the vast amount of sequence space sampled in the human microbiome (167, 168), efforts to inventory MA genes computationally is impeded by the high amino acid sequence similarity within family (e.g. GrdI (34)), the many superfamily members with unknown functions (CutC (27)), and the presence of a stop codon that encodes pyrrolysine (MttB (11)). At the gene level we validated GAMAdb

inclusion using bioinformatic analyses to identify clusters of enzymes with shared quaternary amine biochemistry from their close-functionally disparate neighbors (i.e. CutC from other GRE (27)) (Figure 4.10). Additionally, for the proatherogenic genes we confirmed the active sites and assigned substrates for these that are known (Methods). An exception is the enzymes for carnitine (CntA) and butyrobetaine (YeaW) which we report together as specificity cannot be inferred from sequence information (22). In total, our well curated GAMA-gene database included 5,374 (1,597 unique) proatherogenic microbial genes, with *cutC* (choline), *cntA/yeaW* (carnitine/butyrobetaine), and *grdI* (glycine betaine) accounting for 25, 16, and 12 percent of the entire GAMA-gene database (Figure 4.9B). For each gene, this represents 3 to 12 fold more unique genes than was previously reported in prior studies of each individual gene (30, 45, 162) (Methods), demonstrating that human gut microbial communities harbor vast numbers of enzymes that convert quaternary amines from the diet into disease activating TMA.

Unlike the proatherogenic genes, the substrates of the nonatherogenic MttB superfamily genes can only be inferred if the gene contains pyrrolysine, as these are assumed to be trimethylamine specific (28). For the remaining MttB superfamily sequences that could not be assigned a specific quaternary amine substrate, we denoted these as MtxB to indicate an unassigned substrate "X", nomenclature consistent with the MttB superfamily (e.g. MtgB for glycine betaine (28), MtcB for carnitine (25)). While not acting directly on TMA, we included methyltransferase genes specific for dimethylamine (DMA, n=61) and monomethylamine (MMA, n=70), as methanogens in the gut convert TMA to these nonatherogenic products (10, 11). In summary, the GAMA-gene database included 3,022 nonatherogenic genes (1434 unique) and was composed of 15% pyrrolysine *mttB*, 28% non-Pyrrolysine *mttB* (here denoted *mtxB*), and 4% methyltransferases for DMA and MMA in the entire GAMA-gene database. The

specificity of our analyses to the gut samples, coupled with the comprehensiveness afforded by sampling genomes independent of cultivation, ensure GAMAdb enables new avenues of research into microbial mechanisms for ameliorating cardiovascular disease risk.

Of particular interest was the unprecedented sampling of nonatherogenic *mttB* superfamily gene diversity (Figure 4.9C). For context, the GAMA *mttB* gene content is 2.5 fold more than the only prior report (28), which was restricted to complete genomes, from a subset of cultivated organisms. Here we used a sequence similarity network to visualize the gene similarity across this superfamily (n=1,031 nodes), revealing 17 clusters, with the presence or absence of pyrrolysine most impacting cluster assignment. Within clusters, genes were most closely positioned based on vertical descent of their genome host, and not substrate specificity. For example, two experimentally verified non-pyrrolysine methyltransferases from the same *Eubacterium* genome cluster tightly in clade 8, even though one uses carnitine (25) and the other proline betaine (32), another quaternary amine (Figure 4.9C). Our analyses included biochemically characterized representatives from the non-pyrrolysine (n=3) and pyrrolysine containing (n=3) members of the superfamily, which collectively were assigned to two clusters (Figure 4.9C). Nearly 41% of the GAMA newly sampled *mttB* homologs reside outside these two clusters, and while inferred to demethylate quaternary amines, these findings highlight the substantial biochemical functions remaining to be elucidated in this superfamily. We note these yet to be described sequences are included GAMAdb to facilitate our subsequent community expression analyses (metatranscriptome, metaproteome) from *in vivo* and laboratory reactors, with the goal to further illuminate possible functional roles.

4.2.3 The GAMA-genome database uncovers gene diversity and abundance within genomes

From the 238,530 gut derived MAG or isolate genomes from this and prior studies (Figure 4.11A), we discovered that the capacity to utilize MAs for energy generation was encoded by 3% of these gut microbial members, or 6,341 genomes (Appendix D). Broadly, microorganisms capable of MA metabolism were assigned to a single archaeal (all are methanogen associated lineages) and 12 bacterial phyla (Figure 4.9A). Bacterial members of the Proteobacteria and Firmicutes contained the most members, reflective of their representation in the original genome collection, while the phylum Desulfobacterota (containing the genera *Bilophila* and *Desulfovibrio*) had proportionally the most genomes when this sampling bias was accounted for (Figure 4.9A). This genome resolved approach provided the first genetic evidence for this metabolism from previously unrecognized lineages including two phyla (Synergistota, Spirochaetota), but also 79 new genera (25% GAMAdb) distributed across the 13 GAMA phyla (Methods). One of the most powerful aspects of GAMAdb is the underlying genomic context which allows users to (i) compare MA gene distribution along taxonomic lines resolved at the genome level, (ii) interrogate MA metabolism within the metabolic capacity of a genome, and (iii) contextualize GAMAdb content from multi-omic datasets.

Given that 16S rRNA taxonomy, and not genome content, is the common linkage to ACVD risk factors, we first sought to identify lineages that could be clearly assigned to certain MA metabolisms. A comparison of MA gene distribution revealed that at broad taxonomic levels like phylum or order, only several lineages could be classified as containing genomes that were exclusively proatherogenic (Fusobacteriota) or nonatherogenic (all Archaea). However, the associated functionality becomes more resolved at genus level, with most representatives of a single genus containing only genomes with atherogenic or proatherogenic gene types (Figure

4.11B). A notable exception is members of the Desulfovibrionaceae, which contained several genomes from *Bilophila sp.* and *Desulfovibrio piger* that simultaneously contained a nonatherogenic *mtxB* and/or *mttB*, along with an atherogenic *cutC* (Figure 4.11B).

Based on these genus level generalizations, it could be tempting to justify linking 16S rRNA taxonomic identity of certain lineages to ACVD risk. However, our analysis demonstrate some critiques of this interpretation. First, many lineages that contain MA genes by genus level naming are not monophyletic at the genome level. For example, by 16S rRNA analysis, members of the genus *Eubacterium* were assigned to genomes in the Lachnospiraceae, Anaerovoracaceae, and to 4 separate lineages with Eubacteriaceae. Articulating this point, the model MA *E. limosum* (ATCC 8486) (25) is assigned to the genome-resolved genus *Eubacterium*, while the isolated *E. halii* is the genus Eubacterium_E of the Lachnospiraceae. While we expect this issue to be better resolved as 16S rRNA taxonomy becomes more tied to genome phylogeny, our analysis also revealed concerns for phylogenetically coherent lineages. Namely, just a because a genus has representatives in GAMAdb, it must be noted that not all genomes for a genus contain this metabolism. It is our hope that the GAMA-gene and -genome resolved view of this metabolism will enable more precise linkages to ACVD in the future.

The GAMA-genome database also enables us to examine MA metabolism at the holistic, genome level. First, this analysis showed the variability in copy numbers for closely related genomes. For instance, *Bilophilia* is third in the number of genomes in GAMAdb but is second in number of genes contributed to GAMAdb because it encodes 1-5 MA genes. Similarly, members of the genus *Eubacterium* in GAMAdb range from having 7-43 genes, hinting at the utility of this genome in metabolizing MA compounds in the gut. For context, the model isolate *E. limosum* has 43 genes non-pyl *mttB* genes, whereas *E.coli* has only a single *cutC* or *cntA* gene

for MA metabolism. The maintenance of these differences in gene copies may indicate relevance of these metabolisms to the overall cellular energy budget, however expression analyses are needed to validate these inferences.

A comparative metabolic analysis using DRAM broadly revealed the metabolic capacities of these GAMAdb genomes. For the genomes containing genes assigned to the MttB superfamily, we confirmed the presence of the methyl branch of the Wood-Lungdahl, which is required subsequent processing of the methyl-CoM or methyl-tetrahydrofolate produced by demethylation of MA (Figure 4.12). Broad analysis of energy metabolism in GAMAdb genomes indicated that 37% have aerobic respiratory capacity, 49% can respire anaerobically (e.g. nitrate, sulfate), while 51% are inferred to be obligate fermenters (Appendix D, Methods). Of the genomes with aerobic capacity, >75% contain proatherogenic cntA (carnitine) or yeaW (butyrobetaine) that encoding oxygenases active with the indicated substrate (33), hinting at linkages between gut inflammation as a source of reactive oxygen species and multiple disease states (6, 7, 173). Other broad capabilities of these genomes include the universal capacity to produce short chain fatty acids, which are host energy source and hormone signaling molecules vital for gut homeostasis (6, 174).

4.2.4 Whole community analysis enabled by GAMA-genome database shows that potential methylamine utilizers are low-abundant, prevalent members of the human gut

To understand MA metabolism in a community context, we utilized the MAGs we reconstructed from 54 fecal and 5 enrichment metagenomes to obtain the genome relative abundance profiles within the feces of each human subject (Appendix D). Here, we selected fecal donors based on a TMA concentration gradient, with near equal representations of samples of each TMA concentration by quartile (Figure 4.11C). To deeply sample this metabolism, we

sequenced up to 65 Gbp per sample, which is 17-800 times more than the average metagenome study commonly used today which has 0.08-3.8 giga-basepairs (Gbp) per sample. We evaluated the importance of this sequencing in MA gene recovery using a rarefaction analysis to show that 32 Gbp is sufficient to saturate MA gene recovery, but that gene recovery of these critical-health related genes doubles significantly when sequencing is doubled from 4 to 8 Gbp/sample (Figure 4.11D). Beyond sequencing depth, our approach was genome-recovery motivated, using a combination of sequential assemblies and coassemblies to reconstruct genomes across the dataset (Appendix D). Ultimately, our approach *de novo* reconstructed 557 medium or high quality MAGs (324 unique genomes). Reflecting the high quality of our gut MAGs reconstructed here, nearly a third of these genomes were maintained in the GAMA-genome database, a number more than any of the other two compilation metagenome studies, despite their increased number of MAGs included for consideration (Figure 4.11A).

Reconstruction of MAGs provided more than just gene and genome representatives for GAMAdb, as quantification of these organisms revealed the abundance and distribution of genomes containing MA genes in the human gut. Mapping reads from 52 metagenomes uncovered the relative genome abundance profiles within each human fecal sample, and highlighted that genomes encoding MA metabolism were rare members, or less than 0.1% average relative abundance. All 15 unique MAGs with the potential for MA metabolism were ranked from 52-324 out of 324 genomes. In contrast to their low abundance, MAGs with the potential for MA metabolism were cosmopolitan in the human gut, present in every human fecal sample (n=52). For instance, the most abundant MA containing MAG was assigned to the genus *Dorea*, which was ranked as the 52nd most abundant member of the overall gut community with an average relative abundance of 0.3%, yet was present in every human. (Figure 4.11E).

Consistently, two thirds of our MAGs had a 50% occupancy across the sampled cohort, including *Dorea (n=2 MAGs)*, *Clostridium M (n=3 MAGs)*, *Bilophilia*, Oscillospiraceae sp., *Eubacterium*, Anaerovoracaceae sp., and *Intestinibacter*. Of these MAGs, 2 were exclusively proatherogenic, 1 was exclusively nonatherogenic, and 4 had undetermined atherogenic response as they contained both non and pro gene types. Our findings also showed no single winner, as these MAGs often co-occurred in the human sample, hinting at niche partitioning and larger metabolic cross-feeding networks that may exist, but are currently underappreciated in the gut. Because these MAGs are prevalent and often co-occur, this highlights the potential substrate competitions handoffs that may be present in the gut microbial community. Collectively, our analysis reveals that keystone metabolisms which are critical for human health may not be encoded in the most abundant genomes, and thus require deeper sequencing than is typically performed, combined with targeted methods for genome recovery.

4.2.5 Proteomic and metabolic evidence from fecal laboratory reactors reveal an active methylamine degrading network.

To understand the metabolic roles of these disease relevant quaternary amines, we constructed laboratory microcosms using fecal material collected from subject 74. Triplicate anoxic microcosms were periodically amended with 1mM glycine betaine, choline, carnitine, butyrobetaine, or TMA to a final addition of 30.4 μmol of substrate and incubated for 25 days (Figure 4.2, 4.13A, Methods). Samples were collected at the time of inoculation (T_0), 13 days after inoculation (T_1), 20 days after inoculation (T_2), and at the final time point (T_F), with the final time points chosen for paired metabolomic and metaproteomic analysis (Data File S. To obtain genomes relevant to these microcosms, metagenomics was performed on one microcosm from each substrate at the final timepoint, resulting in 78 MAGs that were incorporated into the

genome database described above. Recovery of methylamine genomes enhanced methylamine medium and high quality MAG recovery by 3-fold (Figure 4.14). These data show that these low abundant members of the community that may become prevalent when dosed with dietary relevant concentrations of quaternary amine.

Quaternary amine addition significantly restructured the membership of the microbial communities over time relative to the T_0 microbial communities (Figure 4.15). Specifically, communities fed with quaternary amines cluster together, while TMA amended microbial communities are more similar to the no substrate controls. NMDS analysis of methylamine gene metaproteomic data also revealed that among quaternary amine microcosms, carnitine and butyrobetaine were the most similar to each other, likely reflecting their similar chemical structure (Figure 4.1, 4.15).

To contextualize the use of quaternary amines by the gut microbial community relative to cardiovascular disease, we profiled the metaproteome from each microcosm for methylamine genes with peptide recruitment and corresponding metabolite changes (Figure 4.13B-G). In the choline microcosm, proatherogenic choline-TMA lyase (CutC) dominates with choline being converted to TMA 85% from T_0 to T_F (Figure 4.13C). Interestingly, choline utilization in these microcosms is spread among several organisms, and different replicates have different dominating CutC. Specifically, in one microcosm choline utilization mostly carried out by a novel genus of the Oscillospiraceae (21% of the peptides recruited to MA genes), while in the other two replicates a *Clostridium M* dominates (18-23% of the peptides recruited to MA genes) (Figure 4.13D). For all 3 replicates, a CutC belonging to a novel genus within the Peptococcaceae is the second most abundant (Figure 4.13D).

Opposite of the choline microcosms, carnitine and butyrobetaine microcosms exclusively produce nonatherogenic demethylation products (Figure 4.13FG). Interestingly, metabolite profiles showed that the carnitine microcosms had both norcarnitine (direct demethylation product of carnitine) and dimethylaminobutyrate (indirect demethylation product by way of butyrobetaine). This suggests that carnitine is first being converted to butyrobetaine and then demethylated to dimethylaminobutyrate, which is corroborated by the peptide recruitment to butyrobetaine hydroxylase in the metaproteome. Also, of note is that this carnitine-amended microcosm does not produce TMA. Coupled to lack of CntA peptide recruitment and the loss of carnitine over the 25-day period, we contribute evidence for an anaerobic mechanism for the degradation of carnitine that does not lead to TMA. Both carnitine and butyrobetaine microcosm metaproteomes recruitment to GAMAdb showed that the same non-pyrrolysine containing methyltransferase (MtxB) from *Eubacterium* was the dominant protein.

The glycine betaine-amended microcosms were the only ones to show production of both proatherogenic TMA and nonatherogenic dimethylglycine. For the glycine betaine microcosms, nonatherogenic processes dominate, with 50% of glycine betaine added demethylated to dimethylglycine by a *Eubacterium* non-pyrrolysine containing methyltransferase. TMA is less prevalent, with 14% being cleaved from glycine betaine primarily by *Clostridium M* glycine betaine reductase (GrdI) (Figure 4.13E). The glycine betaine enrichment highlights potential competition in the gut that may occur for this particular quaternary amine.

Batch-operated laboratory microcosms coupled to the methylamine gene database more readily permitted the quantification metabolic by-products generated by the gut microbial consortia (Figure 4.13). Key outcomes included (i) unveiling the methylamine diversity harbored in the human gut (ii) quantification of proatherogenic and nonatherogenic metabolites by quaternary

amine substrate, and (iii) discovering interconnected metabolisms that may be essential to the gut metabolic economy.

4.2.6 Methylamine gene database abundance profiles predict cardiovascular disease in humans

To quantify the relevance of GAMAdb and the corresponding mechanisms uncovered in laboratory microcosms, we analyzed publicly available human gut metagenomic (29), metranscriptomic (175), and metaproteomic (176) data in light of the methylamine gene database (Figure 4.16A-D). Using GAMAdb nucleotide and amino acid sequences, mapping multi-omic data revealed that MA metabolism is harbored in more than 50% of gut samples across three different studies and data types. These genes were detected in every metagenome sample, 82% of metatranscriptomic samples, and 58% of metaproteomic samples surveyed here. Considering each gene type individually revealed that proatherogenic choline-trimethylamine lyase (*cutC*) and nonatherogenic non-pyrrolysine methyltransferases (*mtxB*) were on average most abundant across data types and studies, while nonatherogenic monomethylamine methyltransferase (*mtmB*) genes were the least detected.

We next wanted to apply GAMAdb to a disease model and hypothesized that given the metabolism and previous published work on MA concentrations linked to ACVD that we could use the gut metagenome from a subject to predict that subject does or does not have atherosclerotic ACVD. Using gut metagenomes from a cohort of 218 individuals with atherosclerotic cardiovascular disease and 187 healthy controls (29), reads were mapped to the methylamine gene database to obtain relative abundance of genes across humans with and without cardiovascular disease. Across all subjects, proatherogenic genes were more abundant than nonatherogenic genes, with degradation of carnitine and butyrobetaine (*cntA/yeaW*) having the highest median relative abundance across all gene types (Figure 4.16B). For nonatherogenic

genes, non-Pyrrolysine containing methyltransferases (*mtxB*) were the most abundant. Mapping data also showed that a significant number of people are outliers from the mean, suggesting some gut microbiomes have a disproportionate abundance of these MA genes (Figure 4.16B). Consistent with atherogenic status defined in Figure 4.1A, when MA gene abundance of ACVD patients was compared to non-ACVD controls, all proatherogenic genes (*cntA/yeaW*, *cutC*, and *grdI*) were statistically increased in ACVD patients, while nonatherogenic *mttB* were statistically decreased (Figure 4.16B).

To quantify the relationship of these methylamine genes to cardiovascular disease, we used logistic regression to predict cardiovascular disease status from gene abundance. Receiver operating curves demonstrate that methylamine gene abundance is significantly predictive of cardiovascular disease with a 71% area under the curve value (Figure 4.16E). Comparing to the widely used blood markers (HDL, LDL, triglycerides), paired data show that area under the curve is 81%. While less, surprisingly the microbiome content was within 10% of predictions with traditional blood markers, demonstrating a clear role of the microbiome in ACVD.

To better resolve the genes and specific organisms associated with cardiovascular disease, we ranked the genes contribution to disease prediction. The most predictive gene types were MttB superfamily members, revealing their potentially important role in nonatherogenic metabolite production from quaternary amines or TMA. Single gene predictions using logistic regression showed that individual gene predictions were better than random chance >50%, but were not as good as all MA genes together, highlighting the importance of considering the overall MA metabolic network. The highest single gene prediction was with proatherogenic CntA/YeaW (AUC =.66) (Figure 4.16F). Beyond gene types, to assess the contribution of particular genes linked to MAGs in ACVD gene prediction, we ranked each GAMAdb gene

contribution. Notably, each proatherogenic and nonatherogenic gene type had a representative in the top 20 most important variables. For the predictive nonatherogenic genes, *mttB* and *mtxB* from *Bilophila* genomes were key important variables. Important proatherogenic genes included those from the genera Escherichia (*cntA/yeaW*), Klebsiella (*cutC*), Clostridium (*grdI*), and Pyramidobacter (*grdI*), as well as families of the Lachnospiraceae and Oscillospiraceae. Collectively, these data highlight the importance of microbial methylamine metabolism in cardiovascular disease and provide potential targets for microbiome based therapeutics.

4.3 Conclusion

The direct link between intestinal microorganisms, dietary quaternary amines, and ACVD risk was established less than 10 years ago (22, 25, 27, 28). While numerous biochemical discoveries have teased apart microbial mechanisms underpinning ACVD pathogenesis, relatively few studies have holistically interrogated this metabolism relative to the entire microbial community. By combining previously published MAGs and isolate genomes, multi-omic informed cultivation methods, and two cohort focused analyses, we have developed GAMAdb, a Gut-Associated Methylated Amine database, that when applied to human gut metagenome datasets, is predictive of ACVD. Our work shows that the majority of known MA metabolism in the human gut is made up of the MttB superfamily. These genes were previously uncharacterized in gut metagenomes and currently have 3 biochemically characterized members. Analysis of these nonatherogenic genes demonstrated an alternate route for quaternary amine degradation, as well as contextualized these genes relative to other MA genes in the gut.

Application of GAMAdb to different datasets and data types revealed several key players in the human gut MA metabolic network including *Eubacterium*, *Bilophilia*, members of the family Anaerovoracaceae, and *Clostridium.* Collectively, our study showed that these organisms

were prevalent across humans, active in human gut metatranscriptomic and metaproteomic datasets, and had peptide recruitment to MA amino acid sequenced when dosed with quaternary amines. Specifically, microcosm experiments coupled to genome resolved metaproteomics was able to untangle the use of quaternary amines by a gut microbial community. Moreover, these anoxic microcosms showed different products for each quaternary amine, with the choline microcosm producing the most TMA, butyrobetaine and carnitine microcosms exclusively producing nonatherogenic by-products, and the glycine betaine microcosms producing both proatherogenic and non-atherogenic by-products. Microcosm TMA yields are consistent with large scale metabolomics studies showing that choline and glycine betaine are predictive of ACVD in humans (5, 7, 159). This study goes beyond previous inventories of MA gene content, by coupling high-resolution multi-omic analysis to cultivation methods, enabling quantification of quaternary amine degradation genome-level processes.

Overall, this study provides critical evidence in support of an intellectual framework for manipulation of the microbiota to combat ACVD. Predictions of ACVD using GAMAdb show that both proatherogenic and nonatherogenic genes are the most important variables. Moreover, communities dosed with quaternary amines have the potential for nonatherogenic demethylation and proatherogenic degradation to TMA, indicating gut communities support both sides of MA metabolism. Coupled to the diversity of these genes outlined by GAMAdb, this study provides targets for microbiota based therapeutics. The consistency of our main findings and their high reproducibility among data generated here as well as other published studies suggests that microbiome-modulating strategies based on MA metabolism could be successfully applied on a population-wide basis.

4.4 Materials and Methods

4.4.1 Study and Data Overview

The current study considered samples collected from 125 individuals aged 21 years or older under the auspices of Dr. Alan George Smulian either at the University of Cincinnati College of Medicine or the University of Cincinnati Medical Center Holmes Hospital Outpatient Services. Each individual provided self-collected fecal and urine samples, along with data on medical history (e.g. antibiotic usage, recent colonoscopy), weight, age, dietary habits, and smoking status (Appendix D). Donor identities were stripped from the paired samples and their associated data and each donor assigned a unique identification number. Fecal 16S rRNA amplicon sequencing data were generated for the total population, while 54 samples were selected based for fecal metagenomic sequencing. Targeted metabolomic analyses of methylated amines (MAs) were carried out on both fecal and urine samples from all 125 individuals. Based on surveys, subjects and their corresponding samples were removed from analyses due to antibiotic use in the last 6 months, lack of patient information, or a colonoscopy in the last 6 months, confining the cohort to 113 subjects. Five sets of donated samples were removed from analyses due to donor antibiotic use and seven were removed for lack of donor de-identified data. Written, informed consent was obtained from all study participants, and subject treatment and experiments with donated samples were approved by Institutional Review Boards of the University of Cincinnati and the Ohio State University.

In this study, a single fecal sample from subject 74 was used to build microcosms to assess microbial interactions among guts microorganisms. The microcosm experiment consisted of six treatments all set up with fecal material from subject 74: (*i*) TMA and fecal material, (*ii*) no substrate and fecal material, (*iii*) glycine betaine and fecal material, (*iv*) carnitine and fecal

material, (*v*) butyrobetaine and fecal material, and (*vi*) choline and fecal material. Each treatment was done in triplicate and consisted of 10% anoxic, fecal slurry (10% by weight) and 90% sterile basal bicarbonate-buffered medium dispensed in Balch tubes sealed with butyl rubber stoppers and aluminum crimps under an atmosphere of N_2/CO_2 [80:20 (vol/vol)]. Before mixing with fecal slurry, the medium (per liter) included 0.25 g ammonium chloride, 0.60 g sodium phosphate, 0.10 potassium chloride, 2.5 g sodium bicarbonate, 10 ml DL-vitamin mixture, and 10 ml DL-mineral mixture and was brought to a pH of 7.0 using 1 mM NaOH (14). Tubes were incubated at 37°C. Samples for metagenomics and metaproteomics were taken at the final (T_F) timepoint, while 16S rRNA and metabolite samples were taken throughout the course of the 25 day incubation (Figure 4.2, 4.13).

4.4.2 Fecal and urine metabolite analyses

Prior to analysis, urine samples were thawed on ice followed by centrifugation of an aliquot at 16100x g for 10 min in a microfuge kept at 4°C. Subsequently 467 µL of each supernatant was mixed with 52.5 µL of D_2O and 5.3 ul of 10 mM Sodium trimethylsilylpropanesulfonate (DSS) (Sigma-Aldrich, St. Louis, MO). The samples were then transferred to 5 x 178 mm NMR tubes.

Fecal samples were removed from the freezer and transferred to a biosafety cabinet on dry ice. A total of 0.2 to 0.5 g (wet weight) of frozen chips of each sample were weighed and transferred to a 5 ml centrifuge tube. To extract metabolites from the fecal samples, 1 ml 0.75 M potassium phosphate buffer (PBS buffer) in 50% D2O, pH 7.2, was added to each tube, resulting either 3x volume/weight dilution (for fecal samples with more than 0.3 g in wet weight) or 5x volume/weight dilution (for fecal samples with less than 0.3 g in wet weight) of the original samples. The slurries were then vortexed for a total of 3 minutes to extract metabolites.

Vortexing was paused several times in order to cool the sample on ice to avoid overheating. The vortexed samples were then centrifuged at 1000x g for 10 minutes at 4°C. The supernatant was transferred to a 1.5 ml microcentrifuge tube and were centrifuged again twice at 4°C (16100x g, 10 min) to remove remaining debris. Total 200 ul of final supernatant were mixed with 100 uM DSS and transferred to a 3 mm x 178 mm NMR tube for NMR analysis.

1D ^{1}H and 2D ^{1}H-^{13}C HSQC NMR spectra were conducted at 298 K on a Bruker Avance III HD 800 MHz (Billerica, MA) at Ohio State campus chemical instrument center (CCIC) NMR facility. Proton NMR, about 4 min for one data set, was acquired using 1.28s acquisition time, 2s relaxation delay, and 64 number of scans. The water suppression was achieved using excitation sculpting with gradients. 2D ^{1}H-^{13}C HSQC was acquired with a standard Bruker pulse sequence in which phase-sensitive was using echo/antiecho-TPPI gradient selection. The experiment parameters include ~4ms acquisition time in ^{13}C dimension, ~80ms acquisition time in ^{1}H dimension, 1s relaxation delay, 16 number of scans, ^{13}C GARP decoupling during acquisition, and data matrix of 2048 X 128. The experimental time is roughly 38 min for one data set. Standards with 100 uM of target metabolites (>98% purify) were analyzed under the same conditions. When appropriate, sample aliquots were spiked with a known concentration of a target metabolite in order to confirm peak assignments.

All NMR data were processed with Bruker Topspin 3.6.1 (Billerica, MA). The data were typically zero-filled one time in both ^{1}H and ^{13}C dimension prior the application of window functions, followed by Fourier transformation, phasing, and baseline correction. Chemical shifts were internally referenced to DSS at 0.00 ppm.

All peak assignments were made based on standards employing commercially available compounds of >98% purity. Existing databases

(https://academic.oup.com/nar/article/46/D1/D608/4616873) and literature reports (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.00730765) were also used to assist peak assignments and metabolite identification. For quantification purposes, integrals of the non-overlapping signal fragments were used. The concentration of a given metabolite were estimated employing standards of known concentration and comparing the integral of peaks to DSS. Urine metabolites were normalized to creatinine while fecal metabolites are reported per g wet weight.

4.4.3 16S rRNA gene sequencing and analysis

Total nucleic acids were extracted using the PowerSoil DNA Isolation kit (MoBio), eluted in 100 µl of elution buffer provided, and stored at −20 °C until sequencing. DNA was submitted for sequencing at Argonne National Lab at the Next Generation Sequencing facility using Illumina MiSeq with 2 × 251 bp paired end reads following established HMP protocols. Briefly, universal primers 515F and 806R were used for PCR amplification of the V4 hypervariable region of 16S rRNA gene using 35 cycles. The 515F primer contained a unique sequence tag to barcode each sample. Both primers contained sequencer adapter regions.

Reads were demultiplexed and analyzed within QIIME2 v.2018.11 (177) using DADA2 (178) to produce an amplicon sequence variant (ASV) by sample table (Appendix D). Raw reads below 10,000 were discarded due to low data quality. The ASV table was analyzed using the statistical package “vegan” (179). Alpha diversity was calculated with the diversity function to investigated both richness and Shannon’s diversity. Beta diversity was calculated by analyzing Bray-Curtis dissimilarities using the relative abundance of samples, and then plotting these values with non-parametric multi-dimensional scaling (NMDS) plots in R. Both a multi response permutation procedure and mean dissimilarity matrix (mrpp) function and an analysis of

similarities (anosim) function were calculated to determine the significance of differences between sample groups.

4.4.4 Fecal metagenomic sequencing, assembly, and binning

Total nucleic acids were extracted from five microcosm samples and 54 human fecal samples using the PowerSoil DNA Isolation kit (MoBio), eluted in 100 μL, and stored at −20 °C until sequencing. DNA was submitted for sequencing at the Genomics Shared Resource facility at The Ohio State University. Libraries were prepared with the Nextera XT Library System in accordance with the manufacturer's instructions. Genomic DNA was sheared by sonication, and fragments were end-repaired. Sequencing adapters were ligated, and library fragments were amplified with five cycles of PCR before solid-phase reversible immobilization size selection, library quantification, and validation. Libraries were sequenced on the Illumina HiSeq 2500 platform, and paired-end reads of 113 cycles were collected. All raw reads from microcosms and fecal samples were trimmed from both the 5′ and 3′ ends with Sickle, and then each sample was assembled individually with IDBA-UD using default parameters (15, 44). Metagenome statistics including amount of sequencing are noted in Appendix D.

All microcosm metagenomes (n=5) and the ten deep sequencing metagenomes (Appendix D) were binned using metabat2 (180) with default parameters. Bins were then assessed for quality using checkM (181) or Amphora . Metagenomic reads from the binned samples were then mapped to bins >50% completion and 10% contamination (medium or high quality bins (134)) 99% identity using bbmap (182). Reads that did not map to medium or high quality bins were then reassembled using IDBA-UD (183), completing iterative assemblies for each of the 15 samples, until no new bins could be recovered (184). The resulting 557 bins were then dereplicated into 324 bins using drep (185). Abundance data reported in Figure 4.11E was based

on the 324 unique bins. Briefly, reads from all 54 metagenomes were mapped to 324 unique bins using bbmap with 90% identity. Read counts were then transformed into relative abundance, accounting for genome length and metagenome size.

4.4.5 GAMAdb construction and analysis

Combining the 557 bins recovered in this study with (i) 700 genomes from isolates in the Human Microbiome Project (HMP) and (ii) 237,273 gut derived metagenome-assembled genomes (MAGs) from previously published studies (167, 168), we obtained 238,530 gut associated genomes for analysis of MA metabolic potential. As outlined in Figure 4.10, each gene type in Figure 4.1A was assessed separately. First, using an experimentally validated amino acid sequence, each gene type was searched against the predicted amino acid sequences of the 238,530 gut associated genomes using BLAST (126), retaining sequences with >60 bitscore. For CutC, CntA, YeaW, and GrdI, sequences were aligned with experimentally validated reference sequences using muscle, and phylogenetic trees were built using RAxML. Individual gene trees were visualized in iTOL, and the branch containing sequences of interest were selected. For the remaining sequences, active residues were confirmed as outlined for CutC (27), CntA, YeaW, and GrdI (34, 36). Of note, is CntA and YeaW, which we report together as specificity cannot be inferred from sequence information alone (22). The remaining sequences with active residues were then incorporated into GAMA gene database, as well as their corresponding genomes into GAMA genome database.

For MttB superfamily genes that do or do not contain pyrrolysine, a different approach was taken due to pyrrolysine interpreted as a stop codon during gene calling (Figure 4.10) (31). After recovery of putative MttB homologs using amino acid BLAST (126), obtained sequences were length filtered to 360 bp and aligned to known MttB superfamily members. Sequences

longer than 360 do not contain pyrrolysine and aligned through the pyrrolysine residue were incorporated into the GAMA gene database as non-pyrrolysine containing MtxB, as well as their corresponding genomes into the GAMA genome database. The remaining truncated genes were then manually called in Geneious (186) from the original genome scaffolds using the amber read-through option to detect pyrrolysine. The resulting sequences that encoded for pyrrolysine were incorporated into the GAMA gene database as pyrrolysine containing MttB, as well as their corresponding genomes into the GAMA genome database.

MttB superfamily genes in GAMAdb were used to construct a sequence similarity network via the EFI-EST webtool (187). Networks were generated with initial edge values of >80%, and sequences with 100% sequence similarity were collapsed into single nodes. The resulting representative node network was visualized with Cytoscape 3.8 (188) using the prefuse force directed layout option and is showcased in Figure 4.9C. Genomes in GAMAdb were analyzed with GTDB-tk (189) for taxonomy, checkM (181) for quality, and DRAM (49) for genome annotation. All tools were ran with default parameters, with results reported in Appendix D.

4.4.6 Microcosm metabolomic data acquisition and analysis

Samples from microcosm experiments were filtered (0.2 μm) at time of collection and sent to the Pacific Northwest National Laboratory for metabolite analysis by NMR. Samples were diluted by 10% (vol/vol) with 5 mM 2,2-dimethyl-2-silapentane-5-sulfonate-d_6 as an internal standard. All NMR spectra were collected using a Varian Direct Drive 600-MHz NMR spectrometer equipped with a 5-mm triple resonance salt-tolerant cold probe. The 1D ^{1}H NMR spectra of all samples were processed, assigned, and analyzed using Chenomx NMR Suite 8.3 with quantification based on spectral intensities relative to the internal standard. Candidate

metabolites present in each of the complex mixtures were determined by matching the chemical shift, J-coupling, and intensity information of experimental NMR signals against the NMR signals of standard metabolites in the Chenomx library. The 1D ^{1}H spectra were collected following standard Chenomx data collection guidelines (143), using a 1D NOESY presaturation (TNNOESY) experiment with 65,536 complex points and at least 512 scans at 298 K. 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. Biological triplicates had similar metabolite pools, with all data reported (Appendix D). Fluid samples from the no-cell control were done in single and showed consistent metabolite concentrations throughout the experiment.

4.4.7 Microcosm metaproteomic extraction, spectral analysis, and data acquisition

Liquid culture (1.2 mL) from each microcosm sample was collected anaerobically, centrifuged for 15 min at 10,000 × *g*, separated from the supernatant, and stored at −80 °C until shipment to Pacific Northwest National Laboratory. Proteins in the pellet were precipitated and washed twice with acetone. Then the pellet was lightly dried under nitrogen. Filter-aided sample preparation kits were used for protein digestion according to the manufacturer's instructions. Resultant peptides were snap-frozen in liquid N_2, digested again overnight, and concentrated to ~30 µL using a SpeedVac (Labconco). Final peptide concentrations were determined using a bicinchoninic acid assay. All mass-spectrometric data were acquired using a Q-Exactive Plus (Thermo Scientific) connected to an nanoACQUITY UPLC M-Class liquid chromatography system (Waters) via in-house 70-cm column packed using Phenomenex Jupiter 3-µm C18 particles and in-house built electrospray apparatus. MS/MS spectra were compared with the predicted protein collections using the search tool MSGF+ (149). Contaminant proteins typically

observed in proteomics experiments were also included in the protein collections searched. The searches were performed using ±20-ppm parent mass tolerance, parent signal isotope correction, partially tryptic enzymatic cleavage rules, and variable oxidation of methionine. In addition, a decoy sequence approach (150) was employed to assess false-discovery rates. Data were collated using an in-house program, imported into a SQL server database, filtered to ~1% false-discovery rate (peptide to spectrum level), and combined at the protein level to provide unique peptide count (per protein) and observation count (that is, spectral count) data. Spectral count data for each identified protein was normalized using normalized spectral abundance frequency (NSAF) calculations, accounting for protein length and proteins per sample (Appendix D). Note that metaproteomics were not done on raw fecal samples. Metaproteomes were mapped to dereplicated GAMAdb predicted amino acid sequences, as well as predicted amino acid sequences of MAGs recovered from this study.

4.4.8 CVD prediction from human gut metagenomic data

All reads were downloaded from EBI from Jie, et al., a study of metagenomes from 218 individuals with atherosclerotic cardiovascular disease (ACVD) and 187 healthy controls (29). Adapters were stripped using bbduk.sh with the parameters ktrim=r, k=23, mink=11, hdist=1. Reads were trimmed using sickle with default parameters. Reads were mapped to GAMAdb genes using bbmap.sh (bbtools suite (182)) using perfectmode=t and ambiguous=random. Counts were extracted from the bbmap covstats output and compiled into a table. The counts were then transformed to geTMMs (190).

GeTMMs were then used in a logistic regression model using scikit-learn (191) to predict ACVD status (0=No ACVD, 1=ACVD) as designated in (29). Models were evaluated using stratified 10 fold cross-validation with mean false positive and true positive rates reported and

used to calculate the area under the receiver operator characteristic curve (AUC-ROC) (192). Additionally models were trained based on gene type where the geTMMs of all genes of the same type were summed per sample and then used in the same model and based on methylating or demethylating. We also used each gene type also to evaluate the model using the summed per sample gene abundances for each gene only as the input to the model. Finally, to compare to the classification of blood marker levels we built a model of ACVD status using triglyceride mmol/L, LDL mmol/L and HDL mmol/L from (29) using the sample model structure.

4.4.9 Transcriptome mapping of published data

All reads were downloaded from EBI from Abu-Ali, et al. (175, 193), a study of metatranscriptomes from adult men. Adapters were stripped using bbduk.sh with the parameters ktrim=r, k=23, mink=11, hdist=1. Reads were trimmed using sickle with default parameters. Reads were mapped to GAMAdb genes using bbmap.sh (bbtools suite (182)) using perfectmode=t and ambiguous=random. Counts were extracted from the bbmap covstats output and compiled into a table. The counts were then transformed to geTMMs (190).

4.4.10 Proteome mapping of published data

All proteome .mgf files were downloaded from Lloyd-Price, et al. (176). Files were then searched against the GAMAdb using MSGF+ (149) using the parameters inst 3, tda 1, ti 1,3, ntt 1 and maxLength 50. After the search files were converted to TSVs using the parameter showDecoy 1. To determine hits, first all hits with a pep q-value greater than removed. Then for each sample proteins with more than one peptide hit were identified. This list of proteins per sample were the ones considered present.

Table 4.1 Overview of methylated amine genes and reactions.

Methylated amine related genes from Figure 4.1A, noting the full reaction and citation (10, 11, 22, 25, 27, 28, 32, 34, 47) for each gene product. Asterisk (*) notes MtxB, which are methyltransferases from the MttB superfamlity capable of quaternary amine demethylation. The substrate specificity for the proteins encoded by three of these genes has recently been published (*mtgB*, *mtpB*, and *mtcB*), demonstrating the potential for quaternary amine demethylation.

GENE	REACTION	PUBLICATION
cutC	choline –> trimethylamine + acetaldehyde	Craciun & Balskus (2012)
grdI	glycine betaine –> trimethylamine + acetate	Meyer, et al. (1995)
cntA	carnitine + oxygen –> trimethylamine + malate	Zhu, et al. (2014)
yeaW	butyrobetaine + oxygen –> trimethylamine + succinate	Koeth, et al. (2014)
mtgB	glycine betaine + cob(I)alamin –> dimethylglycine + methylcobalamin	Ticak, et al. (2014)
mtcB	carnitine + cob(I)alamin –> norcarnitine + methylcobalamin	Kountz, et al. (2020)
mttB	trimethylamine + cob(I)alamin –> dimethylamine + methylcobalamin	Ferguson & Krzycki (1997)
mtbB	dimethylamine + cob(I)alamin –> monomethylamine + methylcobalamin	Paul, et al. (2000)
mtmB	monomethylamine + cob(I)alamin –> ammonium + methylcobalamin	James, et al. (2001)
*mtxB**	quaternary amine + cob(I)alamin –> demethylated quaternary amine + methylcobalamin	Ticak, et al. (2014), Picking, et al. (2019), Kountz, et al. (2020)

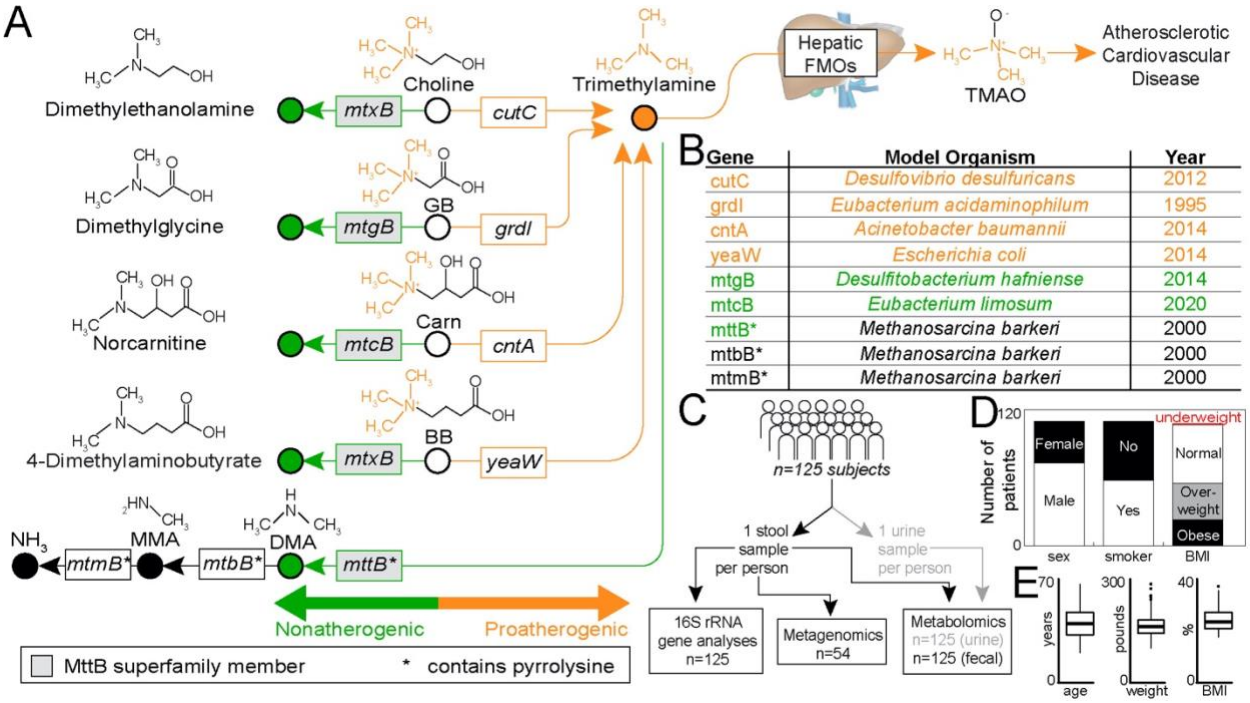

Figure 4.1 Interrogation of methylated amine metabolism in the human gut of 125 subjects using multi-omics analyses.

(A) Overview of microbial methylated amine (MA) metabolism in the human gut leading to proatherogenic (shown in orange) and nonatherogenic (shown in green) metabolites. Genes in the MttB superfamily are shaded in grey, while those encoding a pyrrolysine residue are noted with an asterisk. Abbreviations for metabolites are as follows- GB: Glycine Betaine, Carn: Carnitine, BB: Butyrobetaine, DMA: Dimethylamine, and MMA: Monomethylamine. Genes showcased in this figure are cited in Table 4.1. (B) Table of MA related genes from Figure 4.1A, noting the model organism and year discovered. Asterisk (*) notes genes that encode for pyrrolysine. (C) Outline of analyses performed on 125 human subjects. Twelve subjects were removed from all downstream analysis due to antibiotic use or a colonoscopy in the last six months, constraining the cohort to 113 humans. All further analyses were carried out with the resulting cohort of 113 human subjects. (D) Bar chart showing cohort statistics including sex, smoking status, and BMI category of 113 human subjects. E) Boxplots denote the inner quartile range and median values of age, weight, and BMI statistics across the cohort (n=113). Points above or below boxplots signify outliers, or values outside of the upper or lower quartile.

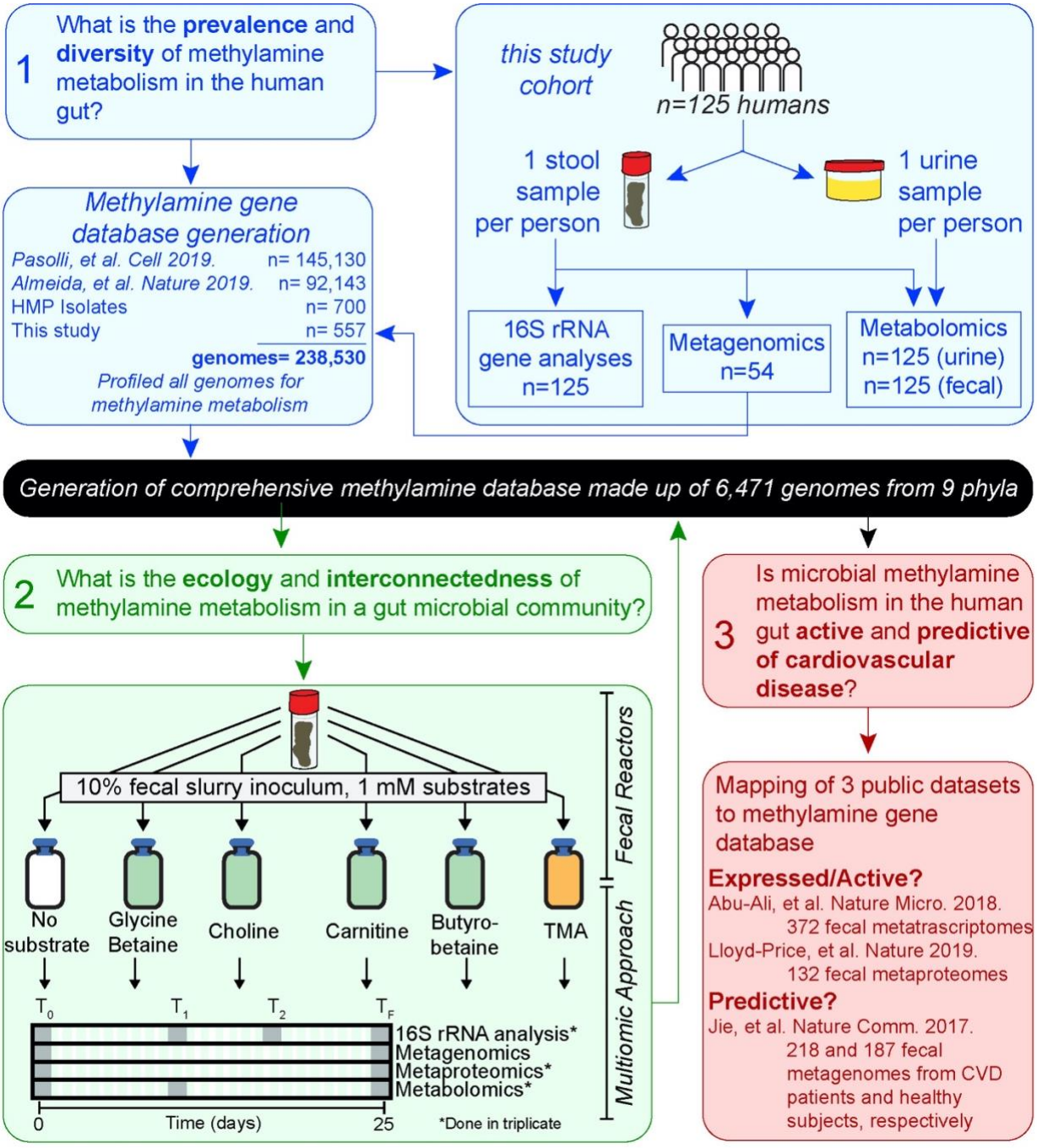

Figure 4.2 Overall experimental design.

Outlines of the objectives for each experiment carried out to interrogate microbial methylamine metabolism in the gut.

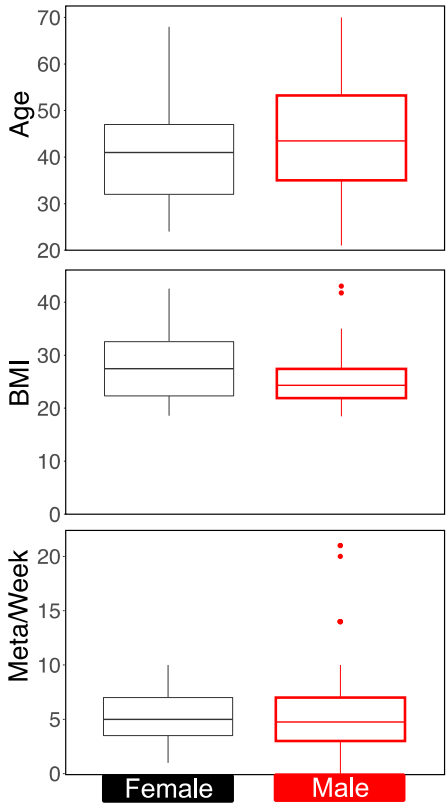

Figure 4.3 Boxplots of human metadata by sex.

Boxes represent the inner quartile range, while points represent outliers.

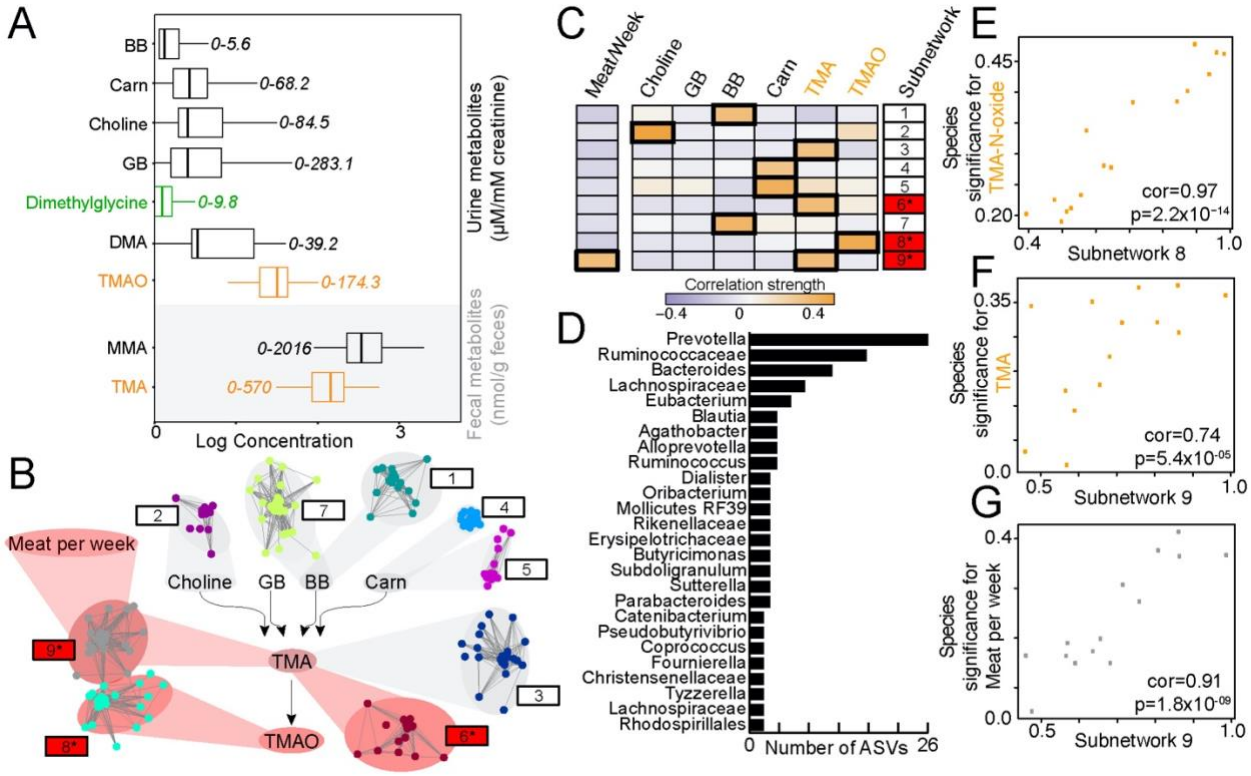

Figure 4.4 Combined community and metabolite analyses reveal microbial subnetworks are predictive of metabolite concentration in the gut.

(A) Boxplots denote the inner quartile range and median log transformed values of fecal and urine MA concentrations across the cohort, with the non-log transformed range of metabolite concentrations stated. Metabolite names and boxplots are colored by proatherogenic (orange) or non-atherogenic (green), while other MAs are black. (B) Weighted gene coexpression network analysis (WGCNA) revealed that 9 modules are significantly correlated to key MA metabolites, labeled 1-9. Colored circles, or nodes, denote features (n=170) within each module (clusters of nodes), with grey lines between features corresponding to intramodule connections. Shading behind modules denotes positive correlations between modules and metabolites. Grey shading notes a significant correlation, while red shading notes module is significantly correlated, as well as predictive of metabolite concentration by Sparse Partial Least Squares (SPLS) analysis. (C) Heatmap shows significant correlations of microbial modules determined by WGCNA to host and metabolite data for a cohort of 113 subjects, with CVD promoting metabolites shown in orange text. Subnetworks that were predictive of host TMA, TMAO, and meat consumption are highlighted by a red box and asterisk (*). Module numbers correspond to module numbers in (Figure 4.4B). (D) Bar chart summarizes the ASV genus level taxonomy of the nine WGCNA modules correlated to methylated amine compounds or meat consumption (Figure 4.4B), with genera that had two or more ASVs present across the nine subnetworks shown. Single microbial subnetworks are strongly associated to TMA-N-oxide (E), TMA (F), and meat consumption (G). The WGCNA approach directly links subnetworks to environmental parameters, i.e. the more the features contribute to the subnetwork structure (topology), the more their abundance are correlated to the parameter. This measure allows to identify subnetworks for which the overall structure, summarized as the eigen vector of the subnetwork, is related to the host metabolites.

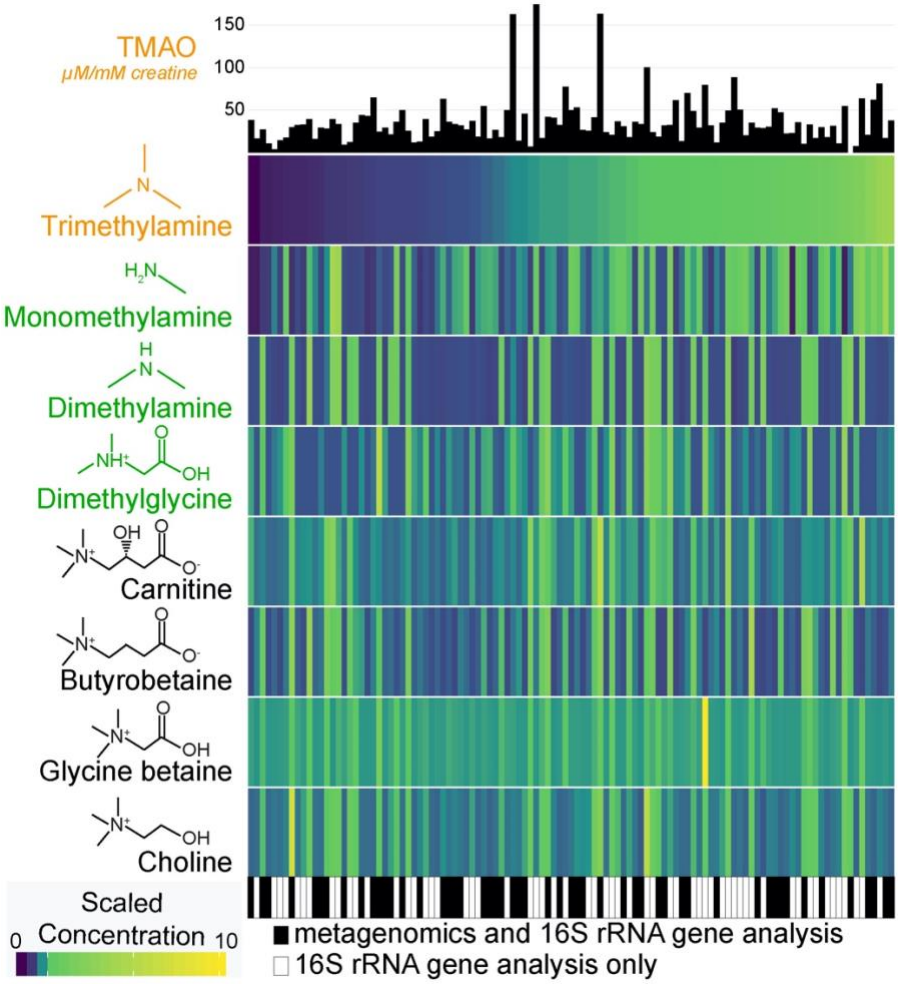

Figure 4.5 Methylated amine concentrations across human subjects.

Heatmap shows the normalized concentration of methylated amine metabolites across urine and fecal samples, with proatherogenic metabolites shown in orange text and nonatherogenic metabolites shown in green. Samples are ordered left to right by increasing TMA concentrations, with fecal samples chosen for metagenomic sequencing noted at the bottom by black and white boxes. Corresponding urine TMAO concentrations are denoted by the black bar chart at the top of the heatmap.

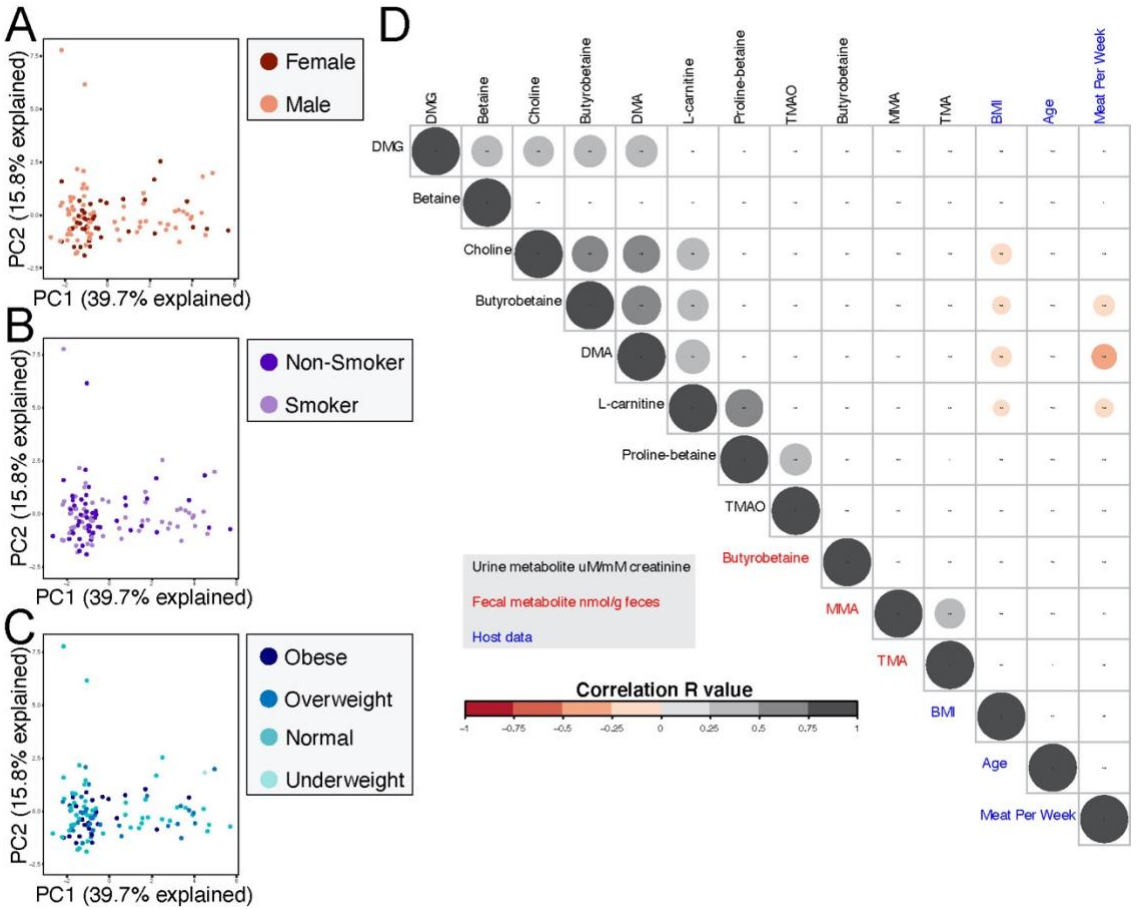

Figure 4.6 Host metadata correlations.

Principle component analysis of methylated amine metabolite concentrations is shown (A-C), with each point colored by (A) Sex, (B) Smoking status, and (C) BMI category. Scatter plots show the first two principle components of the metabolite data, which together account for 55% of the total variation in the metabolite data. (D) Correlation plot shows the only significant correlations (p-value<0.05) among host metadata and metabolite concentrations. Circles are colored and sized by the correlation R value.

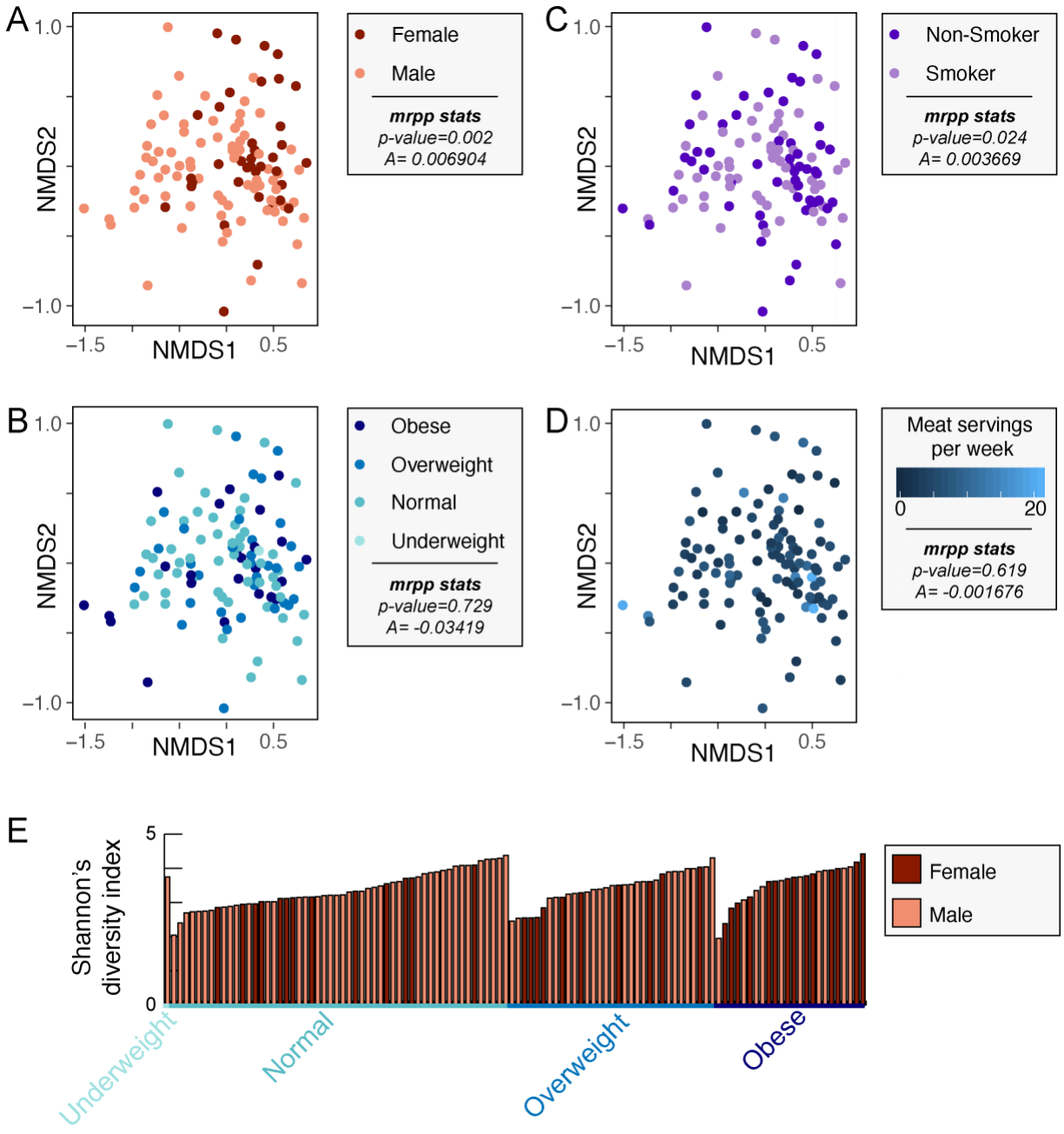

Figure 4.7 Microbial community diversity statistic across the human cohort.

(A-D) The relative similarity of microbial communities across humans was examined by calculating a Bray-Curtis dissimilarity matrix and visualized using non-metric multidimensional scaling (NMDS) in two dimensions. This ordination is colored by several 4 host factors (A-D) and shows that overall structure and membership of the human gut microbial communities examined in this study were not significantly different by sex (A), BMI category (B), smoking status (C), or meat consumption (D). (E) Bar chart shows Shannon's diversity index of cohort (n=113), with each bar representing one gut microbial community. Bars are colored by sex and labeled by BMI category on the x-axis.

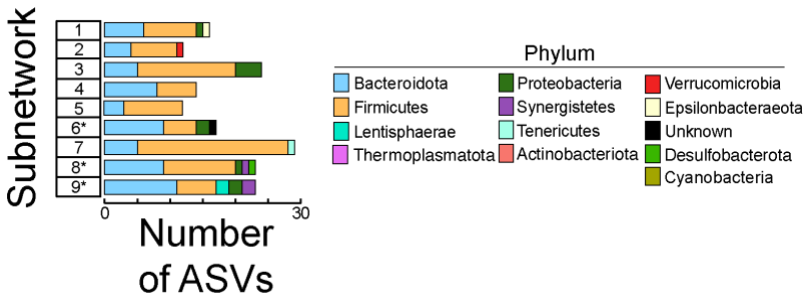

Figure 4.8 Weighted Gene Co-expression Network Analysis (WGCNA) subnetwork membership shown in Figure 4.4.

Stacked bar chart of WGCNA subnetwork membership by Phylum for the 9 subnetworks correlated to methylated amine concentrations.

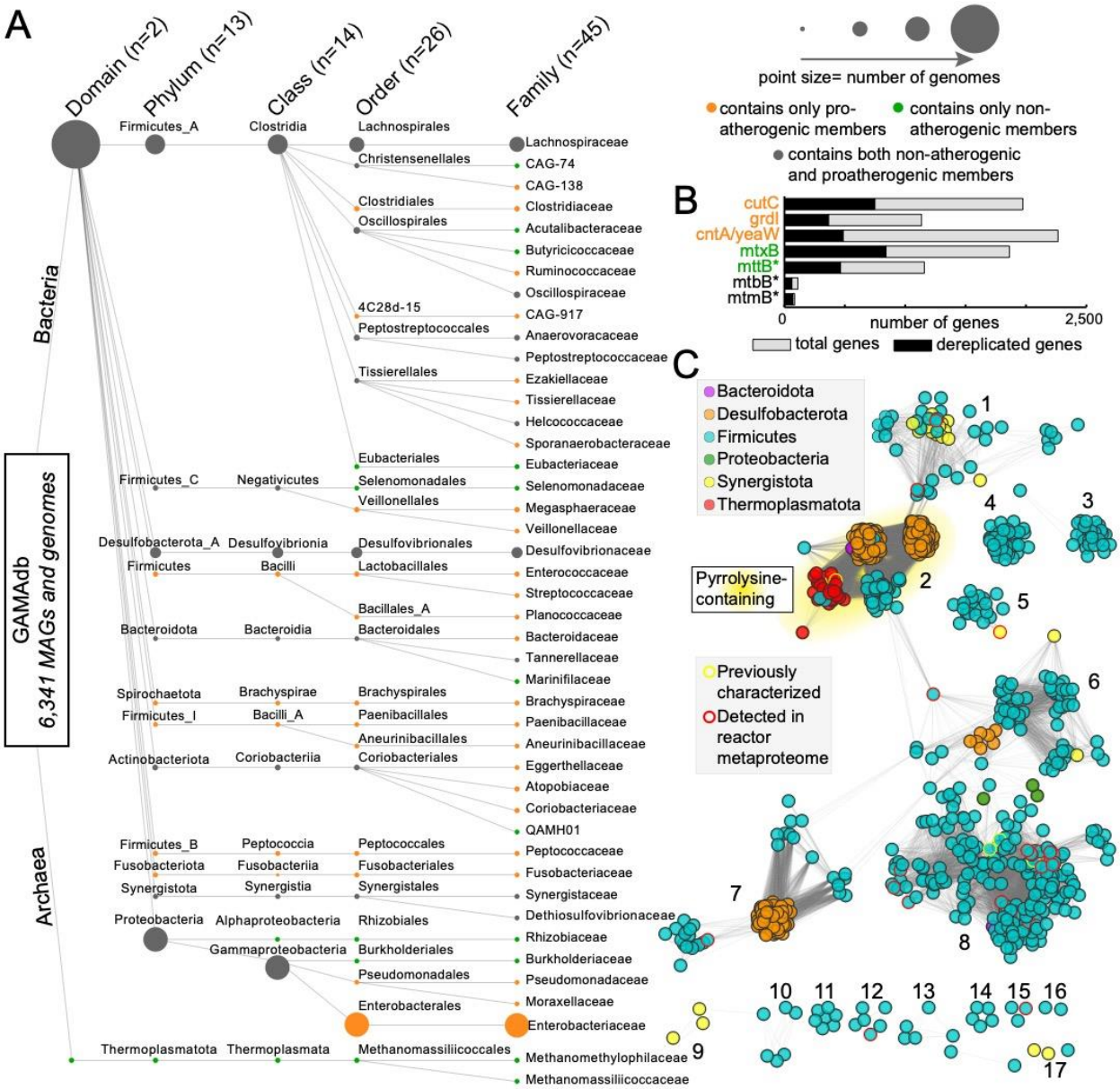

Figure 4.9 Gut- Associated MethylAmine database (GAMAdb) uncovers diversity of MA metabolism in the gut.

(A) Bubble plot shows taxonomy levels defined by GTDB-Tk (left to right: Domain, Phylum, Class, Order, Family) of isolate genomes and MAGs represented in GAMAdb, with bubble size denoting number of genomes and bubble color denoting proatherogenic and nonatherogenic status as outlined in Figure 4.1. (B) Bar chart highlights the number of genes by type in GAMAdb, with bars noting the number of dereplicated genes in black. (C) A sequence similarity network of the MttB superfamily was constructed such that all nodes are connected by an edge if the pairwise sequence similarity is >80% sequence identity. Each of the 1,031 nodes represents one or more genes in GAMAdb, with identical genes (100% amino acid identity) collapsed into single nodes. Nodes are colored by the phylum and outline signifies previously characterized genes (yellow) and detection in reactor metaproteome (red, described below).

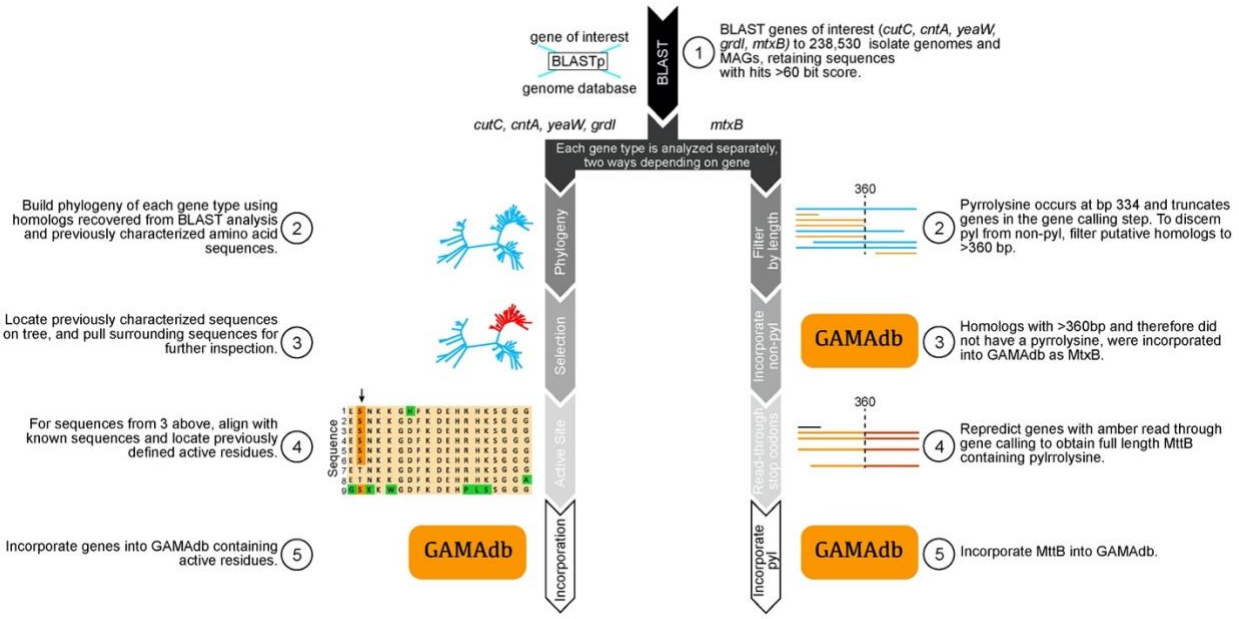

Figure 4.10 Bioinformatic workflow for building GAMAdb**.**

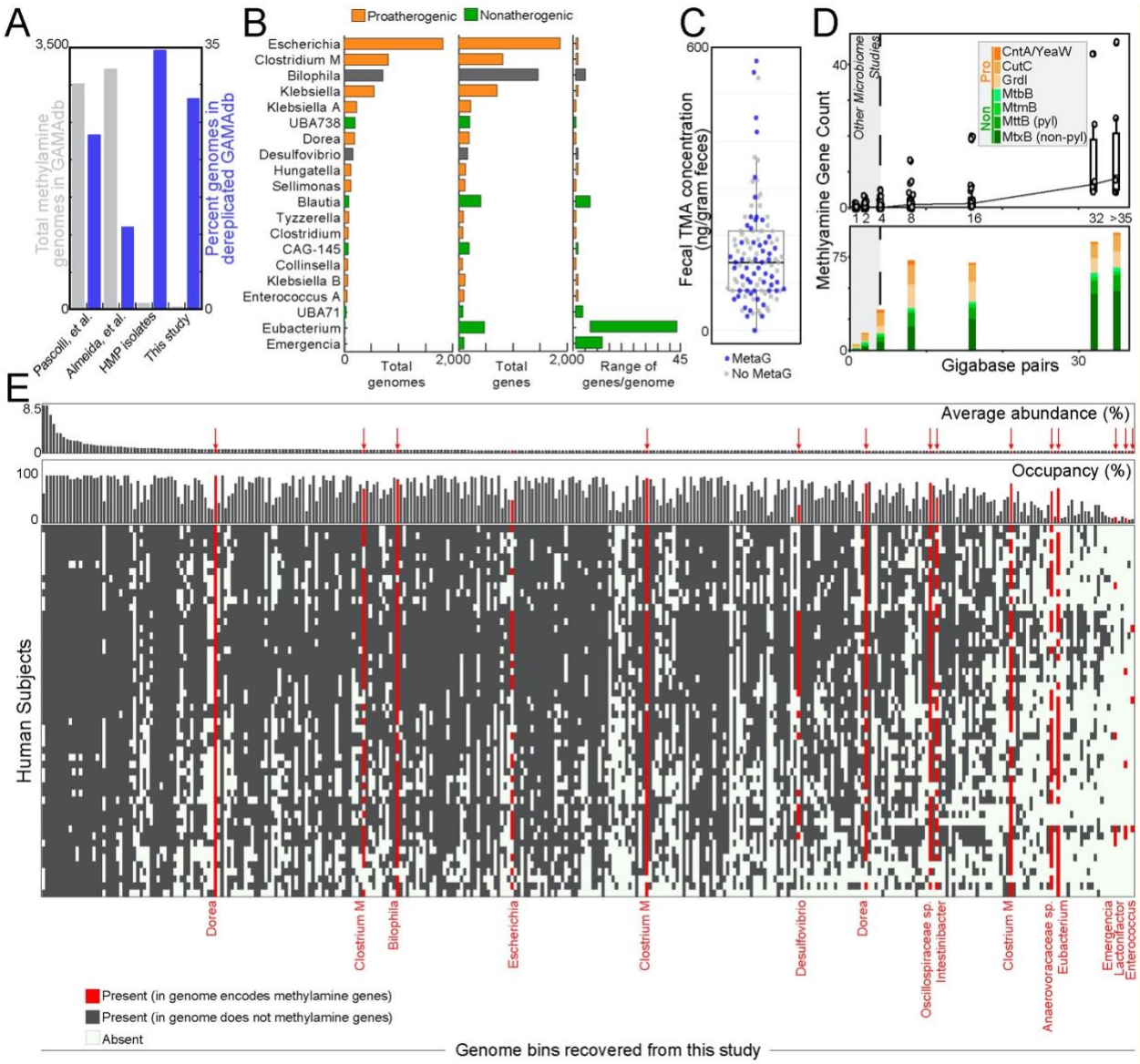

Figure 4.11 Prevalence and abundance of methylated amine utilizers.

(A) Barchart showcases the source of 6,341 genomes encoding MA metabolism in GAMAdb (grey bars) retrieved from a larger set of 238,530 MAGs, as well as the percent unique genomes that remained in GAMAdb (blue bars). (B) Summarized by genus, barcharts note the total number of genomes in GAMAdb, total MA genes, and number of MA genes per genome, with the top 20 represented genera shown. Bars are colored by proatherogenic (orange) or nonatherogenic (green), or both (grey) to show if all members encode a particular gene type as defined in Figure 4.1. (C) Boxplot shows the inner quartile range and median of fecal TMA concentrations across the cohort, with points representing individual subject TMA concentration and colored by metagenomic sequencing of corresponding microbial community. (D) Rarefaction curve (top) and corresponding stacked bar chart (bottom) represent a subsampling of the reads from 54 human gut metagenomes in this study mapped to the methylamine gene database. Methylamine gene count for both graphs represents the number of unique genes

detected in each subsampled dataset. Most recent microbiome studies to date have sequencing depth less than 4Gbp and shows what methylamine genes would be missed at lower sequencing depths. (E) Heatmap shows presence and absence of 325 genomes recovered from each human metagenome (n=52), with genomes containing methylated amine genes highlighted in red and genus specified at the bottom. Bar graphs at the top of the heatmap denote average relative abundance (top) and occupancy (bottom) of each genome across 52 human metagenomes. Genomes across all three graphs are ordered by average relative abundance to show rank of each genome.

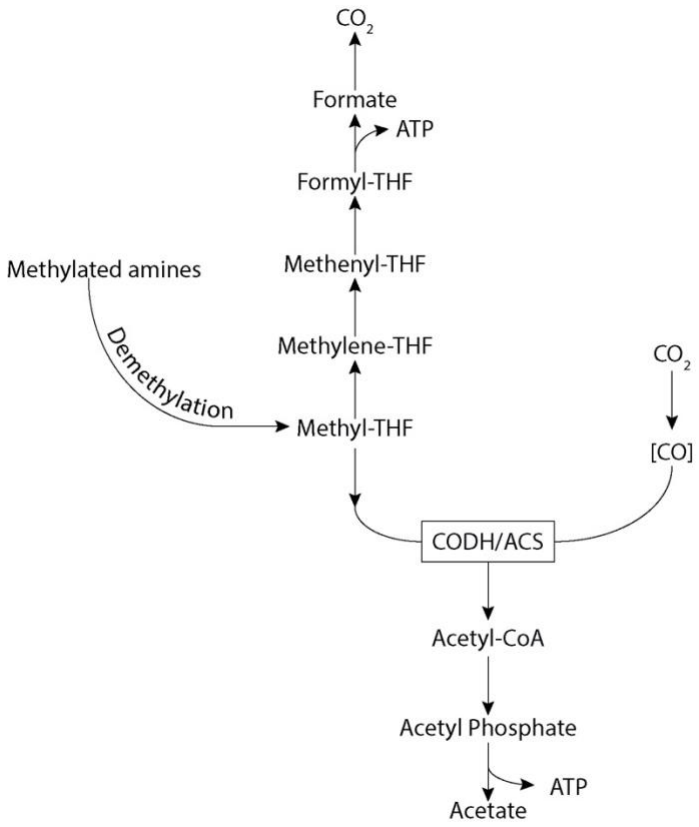

Figure 4.12 Incorporation of demethylation by-product, methyl-THF, into the Wood-Lungdahl pathway (26, 28).

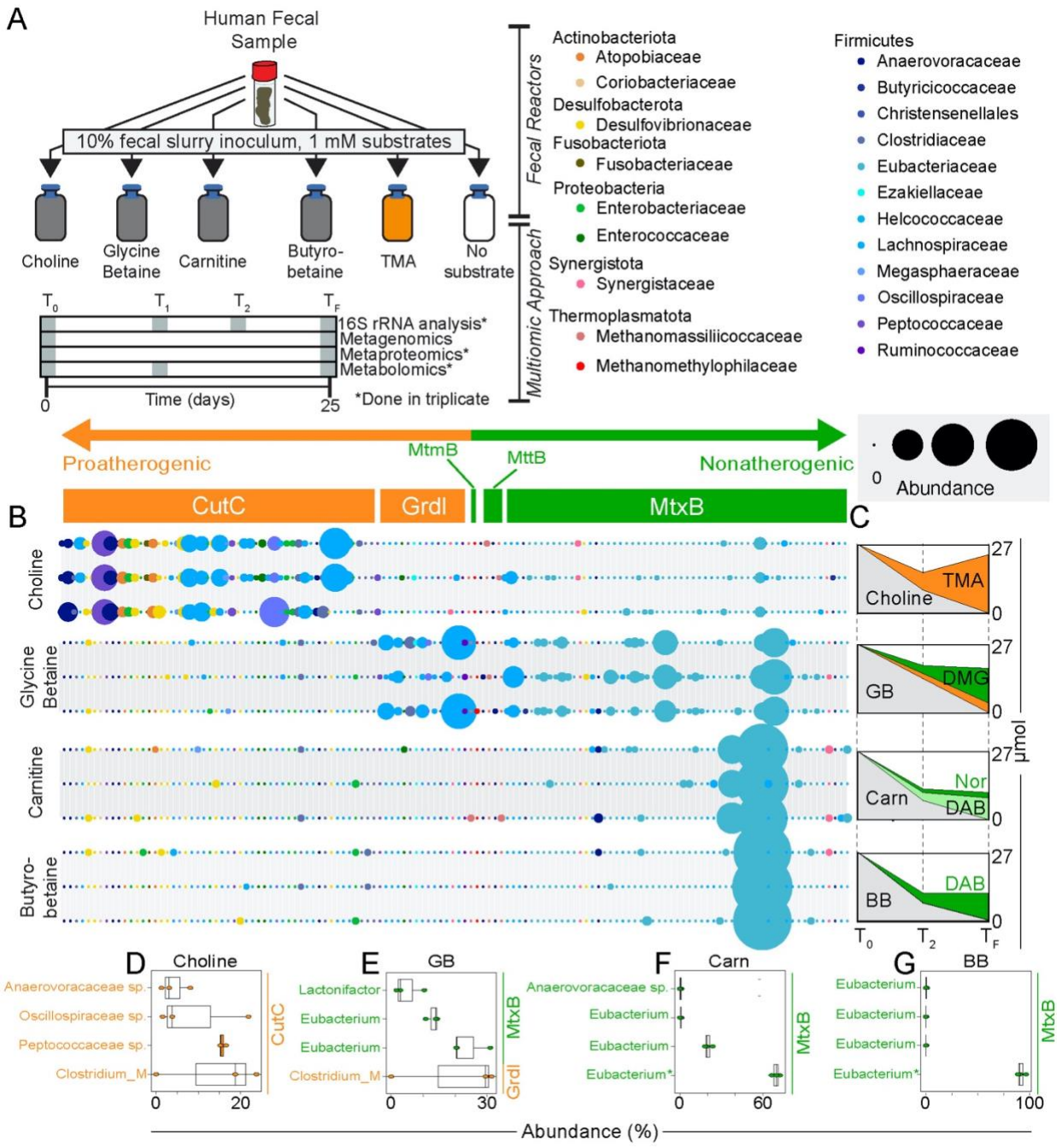

Figure 4.13 Enriched fecal microbial communities degraded methylated amines.

(A) Diagram shows the enrichment experimental design, including treatments and sampling scheme. Using fecal material from Subject 74, 6 enrichment types were started in triplicate and sampled periodically over 25 days. Timeline denotes points at which samples were taken for each multi-omic analysis with grey boxes. (B) Bubble plot displays the relative abundance (circle size) of methylated amine gene entries in GAMAdb determined by peptide recruitment for each replicate of each microcosm using metaproteomics. Bubbles represent one amino acid sequence in the database and are colored by GTDB-Tk family level taxonomy. The full bubble plot including TMA and no substrate controls are shown in Figure S11. (C) Area plots show the

relevant methylamine metabolite concentrations over time in the microcosms, with curve colored by substrate added (grey), proatherogenic metabolite TMA (orange), or nonatherogenic metabolite(s) (green). (D-G) Boxplots show the top four detected GAMAdb entries by metaproteomics for each enrichment, with points and boxplots colored by pro- and non-atherogenic. Boxplots are labeled by genus. Asterisks note a single *Eubacterium* that were active in two separate microcosms.

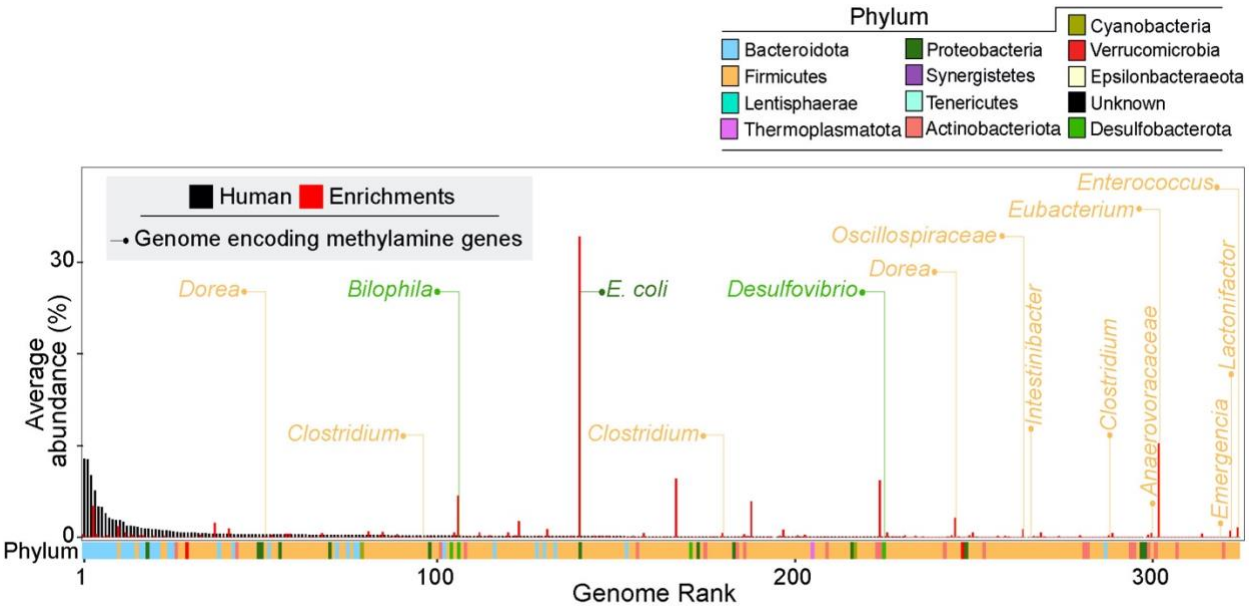

Figure 4.14 Rank abundance curve of all unique MAGs recovered in this study in humans and microcosm experiments.

Rank abundance curve of all 324 unique MAGs recovered in this study shows average relative abundance of each genome across humans in black, with corresponding abundance of the same genomes in the microbial communities enriched with QAs in red. Genome phylum is denoted along the bottom of the rank abundance curve. Genomes with the potential for methylamine metabolism, as defined in Figure 4.1A are highlighted with the genus name and colored by Phylum.

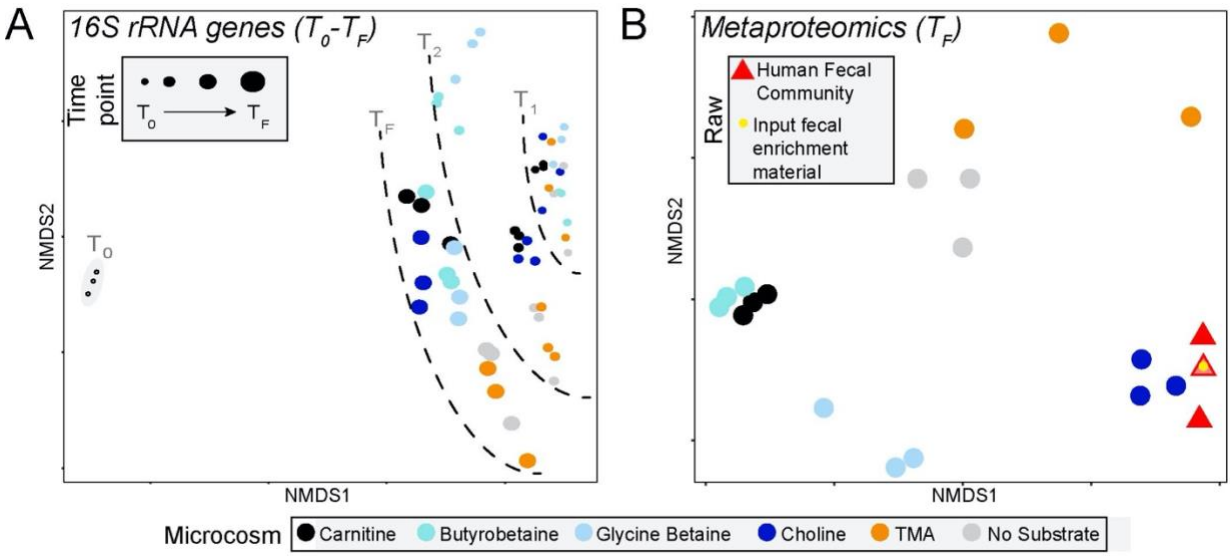

Figure 4.15 Microbial community response to methylamine addition.

(A) NMDS showing the microbial response to different methylamine additions over time, with each point representing one microbial community at a given point in time. Points are sized by timepoint and colored by methylamine addition. (B) NMDS displaying the difference between methylamine gene metaproteomic content among microcosms at the final timepoint.

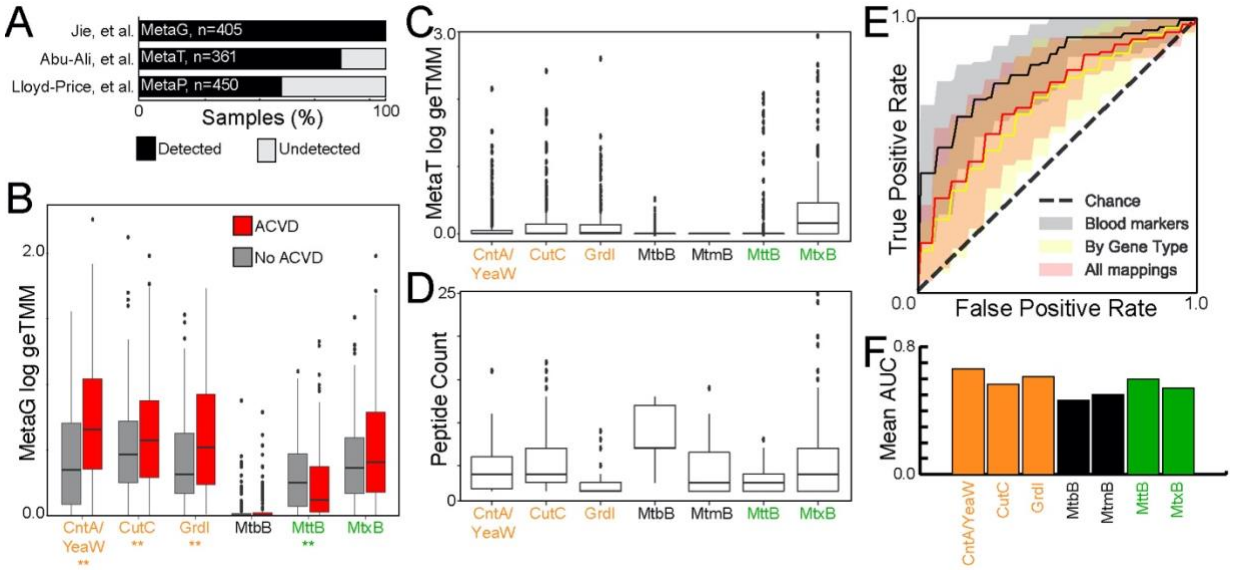

Figure 4.16 GAMAdb genes predict Atherosclerotic CardioVascular Disease (ACVD) in humans.

(A) Bar chart denotes the percentage of samples per study that members of GAMAdb is present or active. Studies include metagenomic data from a cohort of 218 individuals with atherosclerotic cardiovascular disease and 187 healthy controls (29), metatranscriptomic data from 361 adult men (193), and metaproteomic data from longitudinal sampling of 132 patients with irritable bowel syndrome (176). (B) Boxplots show the presence of GAMAdb genes in ACVD patients or non-ACVD patient control metagenomes from Jie, et al. Gene names are colored by proatherogenic (orange) and nonatherogenic (green) as outlined in Figure 4.1A, and boxplots are colored by ACVD (red) or no ACVD (grey). Double asterisk below gene name indicates significant difference of gene abundance between ACVD and non-ACVD controls. (C) Boxplots show the detection of GAMAdb genes in metatranscriptomics from 361 adult men from Abu-Ali, et al. Genes are colored by proatherogenic (orange) and nonatherogenic (green). (D) Boxplots show the detection of GAMAdb genes in metaproteomics from 132 patients with irritable bowel syndrome from Lloyd-Price, et al. (176). Genes are colored by proatherogenic (orange) and nonatherogenic (green). (E) Using gut metagenomes from a cohort of 218 individuals with atherosclerotic cardiovascular disease and 187 healthy controls (Jie, et al.), reads were mapped to the methylamine genome database. ROC curves show the ability of blood markers (LDL, HDL, and triglycerides), gene type (yellow), or each GAMAdb entry mapping (all mappings, red) to predict CVD status in humans. (F) Bar chart shows mean area under curve (AUC) for individual gene predictions.

Chapter 5: Conclusion

The collective aim of this dissertation was to piece together MA metabolism across microbial environments to uncover the diversity and interconnectedness of these metabolic processes, but also understand how the environmental context shaped these metabolisms. By interrogating microbial communities using a suite of multi-omics methods across a range of scales in two ecosystems, this overarching aim was met. In hydraulically fractured shales, microbial abundance patterns of MA cycling microorganisms were able to predict *in situ* metabolite concentrations at the field scale (Chapter 2-3). In the human gut, abundance patterns of MA genes in fecal metagenomes were able to predict cardiovascular disease status in humans (Chapter 4). Chapters 2-4 have defined the MA reactions relevant to hydraulically fractured shales and human guts, knowledge that can be leveraged for genome scale metabolic models that account for MA metabolism.

Beyond the contents in this dissertation, I also contributed my knowledge of MA metabolism to other systems, including wetlands soils (14) and prairie pothole lakes (19). Additionally, further reports by other groups documenting this metabolism have increased since the start of this dissertation, further solidifying the prevalence of this microbial metabolism across ecosystems. To provide this broader ecosystem context of MA metabolism, the following paragraphs summarize other studies in which I or others have interrogated MA metabolism.

Wetland soils are one of the largest natural contributors to methane emissions, yet current biogeochemical models of methanogenesis exclude methylotrophic sources, focusing instead only on hydrogenotrophic and acetoclastic methanogens (194, 195). There is growing body of research that suggests in certain soils and lake systems, methylotrophic metabolism may be

important to overall methane flux (12, 14). To understand the contributions of MAs to methane production in mineral, freshwater, hydric soils, we integrated laboratory and field experiments for methylotrophic methanogenesis in Old Woman Creek (OWC), a temperate freshwater wetland located in Ohio, USA (14). Our multi-omic results demonstrated that methylotrophic methanogens of the family *Methanomassiliicoccaceae* were present and active in a freshwater wetland, with metatranscripts indicating that methanol, not methylamines, was the likely substrate under the conditions measured here (14). However, laboratory experiments indicated the potential for other methanogens to become enriched in response to trimethylamine, revealing the reservoir of methylotrophic methanogenesis harbored in these soils is likely much more vast than we could measure under discrete field scales (14). Consistent with our findings, other studies have also shown the potential for methylotrophic methanogenesis in wetlands through isolation or enrichment studies (196, 197).

Other freshwater terrestrial ecosystems, such as the Prairie Pothole Region of North America, composed of millions of small wetlands also has measurable concentrations of methylated amines in soil porewater (19, 198). To understand the role of MAs in this freshwater terrestrial ecosystem, 18 sediment samples were collected for metagenomics from two adjacent wetlands in the Prairie Pothole Region, which were paired to metabolite analysis confirming the presence of MAs. Genomes recovered from this metagenomic study revealed the potential for MA utilization in both sulfate-reducing bacteria, as well as methanogenic archaea, indicating that C1- compounds like trimethylamine and glycine betaine may play a significant role in high sulfate reduction rates and methane emissions in this ecosystem (19, 198). Collectively, these two studies highlight the how MA-utilization maybe climatically relevant and present in high methane emitting ecosystems. Despite this growing appreciation, relative to acetoclastic and

hydrogenotrophic methanogenesis, methylotrophic methanogenesis is comparatively understudied in freshwater terrestrial ecosystems (199–201). This lack of information may have not yet realized implications for biogeochemical models and highlights the need to further explore these microbial metabolisms in terrestrial ecosystems (30, 194, 195, 202).

Considering other ecosystems, in addition to hydraulically fractured shale (Chapter 2-3), these metabolisms have been shown to be increasingly important in other rock-hosted environments (16, 18, 141). Specifically, stable isotope incubations of shale bed samples from 2km below the seafloor showed that microorganisms were capable of metabolizing methylamines, albeit at extremely slow growth rates (16). Likewise, MAs were shown to be important substrates in Movile Cave, an underground ecosystem located near the coast of the Black Sea, where microorganisms utilized them as carbon, energy, and nitrogen sources (18). These studies highlight the importance of organic nitrogen compounds and how their transformations may facilitate microorganismal growth and maintenance across ecosystems.

In summary, methylated amines (MAs) are exceptionally important microbial metabolites in carbon and nitrogen cycling across ecosystems (5, 7, 8, 14–16, 18, 19, 44, 52). Microbial transformations of these metabolites can contribute to the production of greenhouse gasses in terrestrial ecosystems (12–14, 52), allow for persistence in saline and rock-hosted ecosystems (15, 16, 18, 141), and even modulate cardiovascular disease in humans (7, 159). The novelty and complexity of annotating MA microbial metabolisms (described in detail in Chapter 1) have historically impaired the elucidation of these pathways from multi-omic data. While biochemical investigations and isolate characterizations have confirmed the mechanisms for these processes, genome-resolved approaches describing microbial metabolic potential and activity are pushing the boundaries of these yet to be realized microbial metabolisms across ecosystems.

References

1. Ashraf M, Foolad MR. 2007. Roles of glycine betaine and proline in improving plant abiotic stress resistance. Environ Exp Bot 59:206–216.
2. You L, Song Q, Wu Y, Li S, Jiang C, Chang L, Yang X, Zhang J. 2019. Accumulation of glycine betaine in transplastomic potato plants expressing choline oxidase confers improved drought tolerance. Planta 249:1963–1975.
3. Zhalnina K, Louie KB, Hao Z, Mansoori N, da Rocha UN, Shi S, Cho H, Karaoz U, Loqué D, Bowen BP, others. 2018. Dynamic root exudate chemistry and microbial substrate preferences drive patterns in rhizosphere microbial community assembly. Nat Microbiol 3:470–480.
4. Blusztajn JK, Slack BE, Mellott TJ. 2017. Neuroprotective actions of dietary choline. Nutrients 9:815.
5. Tang WHW, Wang Z, Levison BS, Koeth RA, Britt EB, Fu X, Wu Y, Hazen SL. 2013. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. N Engl J Med 368:1575–1584.
6. Brown JM, Hazen SL. 2018. Microbial modulation of cardiovascular disease. Nat Rev Microbiol 16:171–181.
7. Tang WHW, Li DY, Hazen SL. 2019. Dietary metabolism, the gut microbiome, and heart failure. Nat Rev Cardiol 16:137–154.
8. Gibb SW, Mantoura RFC, Liss PS, Barlow RG. 1999. Distributions and biogeochemistries of methylamines and ammonium in the Arabian Sea. Deep Sea Res Part II Top Stud Oceanogr 46:593–615.
9. Oren A. 1990. Formation and breakdown of glycine betaine and trimethylamine in hypersaline environments. Antonie Van Leeuwenhoek 58:291–298.
10. Burke SA, Lo SL, Krzycki JA. 1998. Clustered genes encoding the methyltransferases of methanogenesis from monomethylamine. J Bacteriol 180:3432–3440.
11. Paul L, Ferguson DJ, Krzycki JA. 2000. The trimethylamine methyltransferase gene and multiple dimethylamine methyltransferase genes of Methanosarcina barkeri contain in-frame and read-through amber codons. J Bacteriol 182:2520–2529.
12. Feldewert C, Lang K, Brune A. 2020. The hydrogen threshold of obligately methyl-reducing methanogens. FEMS Microbiol Lett 367:fnaa137.
13. Watkins AJ, Roussel EG, Parkes RJ, Sass H. 2014. Glycine betaine as a direct substrate for methanogens (Methanococcoides spp.). Appl Environ Microbiol 80:289–293.
14. Narrowe AB, Borton MA, Hoyt DW, Smith GJ, Daly RA, Angle JC, Eder EK, Wong AR, Wolfe RA, Pappas A, others. 2019. Uncovering the Diversity and Activity of Methylotrophic Methanogens in Freshwater Wetland Soils. Msystems 4.
15. Daly RA, Borton MA, Wilkins MJ, Hoyt DW, Kountz DJ, Wolfe RA, Welch SA, Marcus DN, Trexler R V, MacRae JD, others. 2016. Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. Nat Microbiol 1:16146.
16. Trembath-Reichert E, Morono Y, Ijiri A, Hoshino T, Dawson KS, Inagaki F, Orphan VJ. 2017. Methyl-compound use and slow growth characterize microbial life in 2-km-deep subseafloor coal and shale beds. Proc Natl Acad Sci 114:E9206--E9215.
17. Cono V La, Arcadi E, Spada G La, Barreca D, Laganà G, Bellocco E, Catalfamo M,

Smedile F, Messina E, Giuliano L, others. 2015. A three-component microbial consortium from deep-sea salt-saturated anoxic lake Thetis links anaerobic glycine betaine degradation with methanogenesis. Microorganisms 3:500–517.
18. Wischer D, Kumaresan D, Johnston A, El Khawand M, Stephenson J, Hillebrand-Voiculescu AM, Chen Y, Murrell JC. 2015. Bacterial metabolism of methylated amines and identification of novel methylotrophs in Movile Cave. ISME J 9:195–206.
19. Martins PD, Danczak RE, Roux S, Frank J, Borton MA, Wolfe RA, Burris MN, Wilkins MJ. 2018. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. Microbiome 6:1–17.
20. Oren A. 1999. Bioenergetic aspects of halophilism. Microbiol Mol Biol Rev 63:334–348.
21. Oren A. 2011. Thermodynamic limits to microbial life at high salt concentrations. Environ Microbiol 13:1908–1923.
22. Zhu Y, Jameson E, Crosatti M, Schäfer H, Rajakumar K, Bugg TDH, Chen Y. 2014. Carnitine metabolism to trimethylamine by an unusual Rieske-type oxygenase from human microbiota. Proc Natl Acad Sci 111:4268–4273.
23. Fonknechten N, Chaussonnerie S, Tricot S, Lajus A, Andreesen JR, Perchat N, Pelletier E, Gouyvenoux M, Barbe V, Salanoubat M, others. 2010. Clostridium sticklandii, a specialist in amino acid degradation: revisiting its metabolism through its genome sequence. BMC Genomics 11:1–12.
24. Schwartz AC, Schäfer R. 1973. New amino acids, and heterocyclic compounds participating in the Stickland reaction of Clostridium sticklandii. Arch Mikrobiol 93:267–276.
25. Kountz DJ, Behrman EJ, Zhang L, Krzycki JA. 2020. MtcB, a member of the MttB superfamily from the human gut acetogen Eubacterium limosum, is a cobalamin-dependent carnitine demethylase. J Biol Chem jbc--RA120.
26. Lechtenfeld M, Heine J, Sameith J, Kremp F, Müller V. 2018. Glycine betaine metabolism in the acetogenic bacterium Acetobacterium woodii. Environ Microbiol 20:4512–4525.
27. Craciun S, Balskus EP. 2012. Microbial conversion of choline to trimethylamine requires a glycyl radical enzyme. Proc Natl Acad Sci 109:21307–21312.
28. Ticak T, Kountz DJ, Girosky KE, Krzycki JA, Ferguson DJ. 2014. A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. Proc Natl Acad Sci 111:E4668--E4676.
29. Jie Z, Xia H, Zhong S-L, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, others. 2017. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 8:1–12.
30. Borrel G, McCann A, Deane J, Neto MC, Lynch DB, Brugère J-F, O'Toole PW. 2017. Genomics and metagenomics of trimethylamine-utilizing Archaea in the human gut microbiome. ISME J 11:2059–2074.
31. Srinivasan G, James CM, Krzycki JA. 2002. Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science (80-) 296:1459–1462.
32. Picking JW, Behrman EJ, Zhang L, Krzycki JA. 2019. MtpB, a member of the MttB superfamily from the human intestinal acetogen Eubacterium limosum, catalyzes proline betaine demethylation. J Biol Chem 294:13697–13707.
33. Boxhammer S, Glaser S, Kühl A, Wagner AK, Schmidt CL. 2008. Characterization of the recombinant Rieske [2Fe--2S] proteins HcaC and YeaW from E. coli. Biometals 21:459–467.

34. Meyer M, Granderath K, Andreesen JR. 1995. Purification and characterization of protein PB of betaine reductase and its relationship to the corresponding proteins glycine reductase and sarcosine reductase from Eubacterium acidaminophilum. Eur J Biochem 234:184–191.
35. Andreesen JR. 2004. Glycine reductase mechanism. Curr Opin Chem Biol 8:454–461.
36. Andreesen JR. 1994. Glycine metabolism in anaerobes. Antonie Van Leeuwenhoek 66:223–237.
37. Barker HA. 1981. Amino acid degradation by anaerobic bacteria. Annu Rev Biochem 50:23–40.
38. Fischbach MA, Sonnenburg JL. 2011. Eating for two: how metabolism establishes interspecies interactions in the gut. Cell Host Microbe 10:336–347.
39. Zindel U, Freudenberg W, Rieth M, Andreesen JR, Schnell J, Widdel F. 1988. Eubacterium acidaminophilum sp. nov., a versatile amino acid-degrading anaerobe producing or utilizing H 2 or formate. Arch Microbiol 150:254–266.
40. Nyyssölä A, Reinikainen T, Leisola M. 2001. Characterization of Glycine SarcosineN-Methyltransferase and Sarcosine DimethylglycineN-Methyltransferase. Appl Environ Microbiol 67:2044–2050.
41. Lai MC, Gunsalus RP. 1992. Glycine betaine and potassium ion are the major compatible solutes in the extremely halophilic methanogen Methanohalophilus strain Z7302. J Bacteriol 174:7474–7477.
42. Cánovas D, Vargas C, Csonka LN, Ventosa A, Nieto JJ. 1996. Osmoprotectants in Halomonas elongata: high-affinity betaine transport system and choline-betaine pathway. J Bacteriol 178:7221–7226.
43. Rebouche CJ, Seim H. 1998. Carnitine metabolism and its regulation in microorganisms and mammals. Annu Rev Nutr 18:39–61.
44. Borton MA, Hoyt DW, Roux S, Daly RA, Welch SA, Nicora CD, Purvine S, Eder EK, Hanson AJ, Sheets JM, others. 2018. Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales. Proc Natl Acad Sci 115:E6585–E6594.
45. Levin BJ, Huang YY, Peck SC, Wei Y, Martinez-Del Campo A, Marks JA, Franzosa EA, Huttenhower C, Balskus EP. 2017. A prominent glycyl radical enzyme in human gut microbiomes metabolizes trans-4-hydroxy-l-proline. Science (80-) 355.
46. Ticak T, Hariraju D, Arcelay MB, Arivett BA, Fiester SE, Ferguson DJ. 2015. Isolation and characterization of a tetramethylammonium-degrading Methanococcoides strain and a novel glycine betaine-utilizing Methanolobus strain. Arch Microbiol 197:197–209.
47. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L, others. 2013. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 19:576–585.
48. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.
49. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodr\'\iguez-Ramos J, Bolduc B, others. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res gkaa621.
50. Dong X, Strous M. 2019. An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs. Front Genet 10:999.
51. Niehaus TD, Thamm AMK, de Crécy-Lagard V, Hanson AD. 2015. Proteins of unknown

biochemical function: a persistent problem and a roadmap to help overcome it. Plant Physiol 169:1436–1442.
52. Borton MA, Daly RA, O'Banion B, Hoyt DW, Marcus DN, Welch S, Hastings SS, Meulia T, Wolfe RA, Booker AE, others. 2018. Comparative genomics and physiology of the genus Methanohalophilus, a prevalent methanogen in hydraulically fractured shale. Environ Microbiol 20:4596–4611.
53. Authur M, Sageman BB. 1994. Marine Shales: Depositional mechanisms and environments of ancient deposits: Annual Review of Earth and Planetary Science, v. 22.
54. Wignall PB. 1994. Black shales. Oxford University Press, USA.
55. Sageman BB, Murphy AE, Werne JP, Ver Straeten CA, Hollander DJ, Lyons TW. 2003. A tale of shales: the relative roles of production, decomposition, and dilution in the accumulation of organic-rich strata, Middle--Upper Devonian, Appalachian basin. Chem Geol 195:229–273.
56. Kargbo DM, Wilhelm RG, Campbell DJ. 2010. Natural gas plays in the Marcellus Shale: Challenges and potential opportunities. ACS Publications.
57. Kerr RA. 2010. Natural gas from shale bursts onto the scene. American Association for the Advancement of Science.
58. Lash GG, Engelder T. 2009. Tracking the burial and tectonic history of Devonian shale of the Appalachian Basin by analysis of joint intersection style. Geol Soc Am Bull 121:265–277.
59. Mouser PJ, Borton M, Darrah TH, Hartsock A, Wrighton KC. 2016. Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. FEMS Microbiol Ecol 92.
60. Vengosh A, Jackson RB, Warner N, Darrah TH, Kondash A. 2014. A critical review of the risks to water resources from unconventional shale gas development and hydraulic fracturing in the United States. Environ Sci Technol 48:8334–8348.
61. Vikram A, Lipus D, Bibby K. 2014. Produced water exposure alters bacterial response to biocides. Environ Sci Technol 48:13001–13009.
62. Akob DM, Cozzarelli IM, Dunlap DS, Rowan EL, Lorah MM. 2015. Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. Appl Geochemistry 60:116–125.
63. Liang R, Davidova IA, Marks CR, Stamps BW, Harriman BH, Stevenson BS, Duncan KE, Suflita JM. 2016. Metabolic capability of a predominant Halanaerobium sp. in hydraulically fractured gas wells and its implication in pipeline corrosion. Front Microbiol 7:988.
64. Booker AE, Borton MA, Daly RA, Welch SA, Nicora CD, Hoyt DW, Wilson T, Purvine SO, Wolfe RA, Sharma S, others. 2017. Sulfide generation by dominant Halanaerobium microorganisms in hydraulically fractured shales. MSphere 2:e00257–17.
65. Elsner M, Hoelzer K. 2016. Quantitative survey and structural classification of hydraulic fracturing chemicals reported in unconventional gas production. Environ Sci Technol 50:3290–3314.
66. Cokar M, Ford B, Kallos MS, Gates ID. 2013. New gas material balance to quantify biogenic gas generation rates from shallow organic-matter-rich shales. Fuel 104:443–451.
67. Ritter D, Vinson D, Barnhart E, Akob DM, Fields MW, Cunningham AB, Orem W, McIntosh JC. 2015. Enhanced microbial coalbed methane generation: a review of research, commercial activity, and remaining challenges. Int J Coal Geol 146:28–41.
68. Park SY, Liang Y. 2016. Biogenic methane production from coal: A review on recent

research and development on microbially enhanced coalbed methane (MECBM). Fuel 166:258–267.
69. Davis KJ, Lu S, Barnhart EP, Parker AE, Fields MW, Gerlach R. 2018. Type and amount of organic amendments affect enhanced biogenic methane production from coal and microbial community structure. Fuel 211:600–608.
70. Tucker YT, Kotcon J, Mroz T. 2015. Methanogenic archaea in marcellus shale: a possible mechanism for enhanced gas recovery in unconventional shale resources. Environ Sci Technol 49:7048–7055.
71. Lipus D, Vikram A, Ross DE, Bibby K. 2016. Draft genome sequence of Methanohalophilus mahii strain DAL1 reconstructed from a hydraulic fracturing-produced water metagenome. Genome Announc 4.
72. Waldron PJ, Petsch ST, Martini AM, Nüsslein K. 2007. Salinity constraints on subsurface archaeal diversity and methanogenesis in sedimentary rock rich in organic matter. Appl Environ Microbiol 73:4171–4179.
73. Struchtemeyer CG, Davis JP, Elshahed MS. 2011. 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.
74. Davis JP, Struchtemeyer CG, Elshahed MS. 2012. 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.
75. Murali Mohan A, Hartsock A, Bibby KJ, Hammack RW, Vidic RD, Gregory KB. 2013. Microbial community changes in hydraulic fracturing fluids and produced water from shale gas extraction. Environ Sci Technol 47:13141–13150.
76. Murali Mohan A, Hartsock A, Hammack RW, Vidic RD, Gregory KB. 2013. Microbial communities in flowback water impoundments from hydraulic fracturing for recovery of shale gas. FEMS Microbiol Ecol 86:567–580.
77. Fichter J, Wunch K, Moore R, Summer E, Braman S, Holmes P, others. 2012. How hot is too hot for bacteria? A technical study assessing bacterial establishment in downhole drilling, fracturing and stimulation operationsCORROSION 2012.
78. Wuchter C, Banning E, Mincer T, Drenzek NJ, Coolen MJL. 2013. Microbial diversity and methanogenic activity of Antrim Shale formation waters from recently fractured wells. Front Microbiol 4:367.
79. Cluff MA, Hartsock A, MacRae JD, Carter K, Mouser PJ. 2014. Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus shale gas wells. Environ Sci Technol 48:6508–6517.
80. Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630.
81. Baker GC, Smith JJ, Cowan DA. 2003. Review and re-analysis of domain-specific 16S primers. J Microbiol Methods 55:541–555.
82. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, Gies EA, Cheng J-F, Copeland A, Klenk H-P, others. 2016. High-resolution phylogenetic microbial community profiling. ISME J 10:2020–2032.
83. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. Science (80-) 312:1355–1359.
84. Henderson G, Cox F, Ganesh S, Jonker A, Young W, Abecia L, Angarita E, Aravena P,

Arenas GN, Ariza C, others. 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep 5:14567.
85. Angle JC, Morin TH, Solden LM, Narrowe AB, Smith GJ, Borton MA, Rey-Sanchez C, Daly RA, Mirfenderesgi G, Hoyt DW, others. 2017. Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. Nat Commun 8:1567.
86. Paterek, J Robert Smith PH. 1988. Methanohalophilus mahii gen. nov., sp. nov., a methylotrophic halophilic methanogen. Int J Syst Evol Microbiol 38:122–123.
87. Boone DR, Mathrani IM, Liu Y, Menaia JAGF, Mah RA, Boone JE. 1993. Isolation and characterization of Methanohalophilus portucalensis sp. nov. and DNA reassociation study of the genus Methanohalophilus. Int J Syst Evol Microbiol 43:430–437.
88. Katayama T, Yoshioka H, Mochimaru H, Meng X-Y, Muramoto Y, Usami J, Ikeda H, Kamagata Y, Sakata S. 2014. Methanohalophilus levihalophilus sp. nov., a slightly halophilic, methylotrophic methanogen isolated from natural gas-bearing deep aquifers, and emended description of the genus Methanohalophilus. Int J Syst Evol Microbiol 64:2089–2093.
89. Spring S, Scheuner C, Lapidus A, Lucas S, Glavina Del Rio T, Tice H, Copeland A, Cheng J-F, Chen F, Nolan M, others. 2010. The genome sequence of Methanohalophilus mahii SLP T reveals differences in the energy metabolism among members of the Methanosarcinaceae inhabiting freshwater and saline environments. Archaea 2010.
90. Davidova IA, Harmsen HJM, Stams AJM, Belyaev SS, Zehnder AJB. 1997. Taxonomic description of Methanococcoides euhalobius and its transfer to the Methanohalophilus genus. Antonie Van Leeuwenhoek 71:313–318.
91. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542.
92. Nesbø CL, Swithers KS, Dahle H, Haverkamp THA, Birkeland N-K, Sokolova T, Kublanov I, Zhaxybayeva O. 2015. Evidence for extensive gene flow and Thermotoga subpopulations in subsurface and marine environments. ISME J 9:1532–1542.
93. Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol 13:13–27.
94. Li S-J, Hua Z-S, Huang L-N, Li J, Shi S-H, Chen L-X, Kuang J-L, Liu J, Hu M, Shu W-S. 2014. Microbial communities evolve faster in extreme environments. Sci Rep 4:6205.
95. Kassen R. 2002. The experimental evolution of specialists, generalists, and the maintenance of diversity. J Evol Biol 15:173–190.
96. Darrah TH, Jackson RB, Vengosh A, Warner NR, Whyte CJ, Walsh TB, Kondash AJ, Poreda RJ. 2015. The evolution of Devonian hydrocarbon gases in shallow aquifers of the northern Appalachian Basin: Insights from integrating noble gas and hydrocarbon geochemistry. Geochim Cosmochim Acta 170:321–355.
97. Moore MT, Vinson DS, Whyte CJ, Eymold WK, Walsh TB, Darrah TH. 2018. Differentiating between biogenic and thermogenic sources of natural gas in coalbed methane reservoirs from the Illinois Basin using noble gas and hydrocarbon geochemistry. Geol Soc London, Spec Publ 468:151–188.
98. Harkness JS, Darrah TH, Moore MT, Whyte CJ, Mathewson PD, Cook T, Vengosh A. 2017. Naturally occurring versus anthropogenic sources of elevated molybdenum in groundwater: Evidence for Geogenic contamination from Southeast Wisconsin, United

States. Environ Sci Technol 51:12190–12199.
99. Lash G, Loewy S, Engelder T. 2004. Preferential jointing of Upper Devonian black shale, Appalachian Plateau, USA: Evidence supporting hydrocarbon generation as a joint-driving mechanism. Geol Soc London, Spec Publ 231:129–151.
100. Oliver J. 1986. Fluids expelled tectonically from orogenic belts: Their role in hydrocarbon migration and other geologic phenomena. Geology 14:99–102.
101. Evans MA. 1995. Fluid inclusions in veins from the Middle Devonian shales: A record of deformation conditions and fluid evolution in the Appalachian Plateau. Geol Soc Am Bull 107:327–339.
102. Engelder T, Lash GG, Uzcátegui RS. 2009. Joint sets that enhance production from Middle and Upper Devonian gas shales of the Appalachian Basin. Am Assoc Pet Geol Bull 93:857–889.
103. Selley RC. 1998. Elements of petroleum geology. Gulf Professional Publishing.
104. L'Haridon S, Corre E, Guan Y, Vinu M, La Cono V, Yakimov M, Stingl U, Toffin L, Jebbar M. 2018. Complete genome sequence of the halophilic methylotrophic methanogen archaeon Methanohalophilus portucalensis strain FDF-1T. Genome Announc 6.
105. Booker AE, Johnston MD, Daly RA, Wrighton KC, Wilkins MJ. 2017. Draft genome sequences of multiple Frackibacter strains isolated from hydraulically fractured shale environments. Genome Announc 5:e00608–17.
106. Galperin MY, Koonin E V. 2010. From complete genome sequence to 'complete'understanding? Trends Biotechnol 28:398–406.
107. Koppel N, Balskus EP. 2016. Exploring and understanding the biochemical diversity of the human microbiota. Cell Chem Biol 23:18–30.
108. Vigil-Stenman T, Ininbergs K, Bergman B, Ekman M. 2017. High abundance and expression of transposases in bacteria from the Baltic Sea. ISME J 11:2611–2623.
109. Sheppard NF, Glover CVC, Terns RM, Terns MP. 2016. The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. Rna 22:216–224.
110. Flemming H-C, Wingender J. 2010. The biofilm matrix. Nat Rev Microbiol 8:623–633.
111. van Wolferen M, Orell A, Albers S-V. 2018. Archaeal biofilm formation. Nat Rev Microbiol 16:699–713.
112. Angel R, Matthies D, Conrad R. 2011. Activation of methanogenesis in arid biological soil crusts despite the presence of oxygen. PLoS One 6:e20453.
113. Cabello P, Roldan MD, Moreno-Vivian C. 2004. Nitrate reduction and the nitrogen cycle in archaea. Microbiology 150:3527–3546.
114. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, others. 2015. An updated evolutionary classification of CRISPR--Cas systems. Nat Rev Microbiol 13:722–736.
115. Shipman SL, Nivala J, Macklis JD, Church GM. 2016. Molecular recordings by directed CRISPR spacer acquisition. Science (80-) 353.
116. Levin BR. 2010. Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. PLoS Genet 6:e1001171.
117. Gómez P, Buckling A. 2011. Bacteria-phage antagonistic coevolution in soil. Science (80-) 332:106–109.
118. Solden LM, Hoyt DW, Collins WB, Plank JE, Daly RA, Hildebrand E, Beavers TJ, Wolfe

R, Nicora CD, Purvine SO, others. 2017. New roles in hemicellulosic sugar fermentation for the uncultivated Bacteroidetes family BS11. ISME J 11:691.
119. Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–1034.
120. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.
121. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.
122. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res 33:W116--W120.
123. Langdon WB. 2015. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min 8:1.
124. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.
125. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. PeerJ 6:e4320.
126. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
127. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. 2014. ITEP: an integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15:1–11.
128. Van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks, p. 281–295. *In* Bacterial Molecular Networks. Springer.
129. Outlook AE. 2017. US Energy Information Administration, 2017. Source https//www eia gov/outlooks/steo.
130. Lipus D, Vikram A, Ross D, Bain D, Gulliver D, Hammack R, Bibby K. 2017. Predominance and metabolic potential of Halanaerobium spp. in produced water from hydraulically fractured Marcellus shale wells. Appl Environ Microbiol 83.
131. Zhang Y, Yu Z, Zhang H, Thompson IP. 2017. 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.
132. Xiao Z, Samuel M, Tibbles R, Moussa O. 2005. Hydraulic fracturing method. Google Patents.
133. Mohan AM, Bibby KJ, Lipus D, Hammack RW, Gregory KB. 2014. 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.
134. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of

bacteria and archaea. Nat Biotechnol. Nature Publishing Group.
135. Konstantinidis KT, Rosselló-Móra R. 2015. Classifying the uncultivated microbial majority: a place for metagenomic data in the Candidatus proposal. Syst Appl Microbiol 38:223–230.
136. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, others. 2011. Evolution and classification of the CRISPR--Cas systems. Nat Rev Microbiol 9:467–477.
137. Mathrani IM, Boone DR, Mah RA, Fox GE, Lau PP. 1988. Methanohalophilus zhilinae sp. nov., an alkaliphilic, halophilic, methylotrophic methanogen. Int J Syst Evol Microbiol 38:139–142.
138. Watkins AJ, Roussel EG, Webster G, Parkes RJ, Sass H. 2012. Choline and N, N-dimethylethanolamine as direct substrates for methanogens. Appl Environ Microbiol 78:8298–8303.
139. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:1–18.
140. Stadtman TC, White Jr FH. 1954. Tracer studies on ornithine, lysine, and formate metabolism in an amino acid fermenting Clostridium. J Bacteriol 67:651.
141. Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD, Stepanauskas R, Richter M, Kleindienst S, Lenk S, others. 2013. Predominant archaea in marine sediments degrade detrital proteins. Nature 496:215–218.
142. White DA. 1973. The phospholipid composition of mammalian tissues. Form Funct phospholipids 441–482.
143. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. 2006. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem 78:4430–4442.
144. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, others. 2012. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science (80-) 337:1661–1665.
145. Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol 12:1–14.
146. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985.
147. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaqué D, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693.
148. Devisetty UK, Kennedy K, Sarando P, Merchant N, Lyons E. 2016. Bringing your tools to CyVerse discovery environment using Docker. F1000Research 5.
149. Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:5277.
150. Elias JE, Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 4:207–214.
151. Borton MA, Sabag-Daigle A, Wu J, Solden LM, O'Banion BS, Daly RA, Wolfe RA, Gonzalez JF, Wysocki VH, Ahmer BMM, others. 2017. Chemical and pathogen-induced inflammation disrupt the murine intestinal microbiome. Microbiome 5:47.

152. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590--D596.
153. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum JR, others. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532:465–470.
154. Shen H, Huang JZ. 2008. Sparse principal component analysis via regularized low rank matrix approximation. J Multivar Anal 99:1015–1034.
155. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. 2008. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol 7.
156. Chong I-G, Jun C-H. 2005. Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst 78:103–112.
157. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. 2007. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics 8:209.
158. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2017. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res 45:39–53.
159. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, DuGar B, Feldstein AE, Britt EB, Fu X, Chung Y-M, others. 2011. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472:57–63.
160. Zhu W, Gregory JC, Org E, Buffa JA, Gupta N, Wang Z, Li L, Fu X, Wu Y, Mehrabian M, others. 2016. Gut microbial metabolite TMAO enhances platelet hyperreactivity and thrombosis risk. Cell 165:111–124.
161. Warrier M, Shih DM, Burrows AC, Ferguson D, Gromovsky AD, Brown AL, Marshall S, McDaniel A, Schugar RC, Wang Z, others. 2015. The TMAO-generating enzyme flavin monooxygenase 3 is a central regulator of cholesterol balance. Cell Rep 10:326–338.
162. Rath S, Heidrich B, Pieper DH, Vital M. 2017. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. Microbiome 5:1–14.
163. Karlsson FH, Fåk F, Nookaew I, Tremaroli V, Fagerberg B, Petranovic D, Bäckhed F, Nielsen J. 2012. Symptomatic atherosclerosis is associated with an altered gut metagenome. Nat Commun 3:1–8.
164. Zuo K, Liu X, Wang P, Jiao J, Han C, Liu Z, Yin X, Li J, Yang X. 2020. Metagenomic data-mining reveals enrichment of trimethylamine-N-oxide synthesis in gut microbiome in atrial fibrillation patients. BMC Genomics 21:1–9.
165. Falony G, Vieira-Silva S, Raes J. 2015. Microbiology meets Big Data: the case of gut microbiota--derived trimethylamine. Annu Rev Microbiol 69:305–321.
166. Sandek A, Bauditz J, Swidsinski A, Buhner S, Weber-Eibel J, von Haehling S, Schroedl W, Karhausen T, Doehner W, Rauchhaus M, others. 2007. Altered intestinal function in patients with chronic heart failure. J Am Coll Cardiol 50:1561–1569.
167. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 2019. A new genomic blueprint of the human gut microbiota. Nature 568:499.
168. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and

Lifestyle. Cell 176:649-662.
169. Tang WHW, Wang Z, Kennedy DJ, Wu Y, Buffa JA, Agatisa-Boyle B, Li XS, Levison BS, Hazen SL. 2015. Gut microbiota-dependent trimethylamine N-oxide (TMAO) pathway contributes to both development of renal insufficiency and mortality risk in chronic kidney disease. Circ Res 116:448–455.
170. Haley M, Jones K. 2017. Livestock, dairy, and poultry outlook. Econ Res Serv United States Dep Agric.
171. Fryar CD, Kruszan-Moran D, Gu Q, Ogden CL. 2018. Mean body weight, weight, waist circumference, and body mass index among adults: United States, 1999--2000 through 2015--2016.
172. Müller E, Fahlbusch K, Walther R, Gottschalk G. 1981. Formation of N, N-dimethylglycine, acetic acid, and butyric acid from betaine by Eubacterium limosum. Appl Environ Microbiol 42:439–445.
173. Bäumler AJ, Sperandio V. 2016. Interactions between the microbiota and pathogenic bacteria in the gut. Nature 535:85–93.
174. R\’\ios-Covián D, Ruas-Madiedo P, Margolles A, Gueimonde M, de los Reyes-Gavilán CG, Salazar N. 2016. Intestinal short chain fatty acids and their link with diet and human health. Front Microbiol 7:185.
175. Mehta RS, Abu-Ali GS, Drew DA, Lloyd-Price J, Subramanian A, Lochhead P, Joshi AD, Ivey KL, Khalili H, Brown GT, others. 2018. Stability of the human faecal microbiome in a cohort of adult men. Nat Microbiol 3:347–355.
176. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, others. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569:655–662.
177. Boylen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander A, Alm EJ, Arumugam A, Asnicar F, others. 2018. QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Prepr 6:e27295v2.
178. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583.
179. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J Veg Sci 14:927–930.
180. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019:e7359.
181. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.
182. Bushnell B. 2018. BBTools. BBMap.
183. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428.
184. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. 2013. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 1:22.

185. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J 11:2864–2868.
186. Olsen C. Geneious R8: A powerful and comprehensive suite of molecular biology tools.
187. Gerlt JA, Bouvier JT, Davidson DB, Imker HJ, Sadkhin B, Slater DR, Whalen KL. 2015. Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochim Biophys Acta - Proteins Proteomics. Elsevier B.V.
188. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504.
189. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2018. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927.
190. Smid M, van den Braak RRJC, van de Werken HJG, van Riet J, van Galen A, de Weerd V, van der Vlugt-Daane M, Bril SI, Lalmahomed ZS, Kloosterman WP, others. 2018. Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. BMC Bioinformatics 19:1–13.
191. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, others. 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830.
192. Davis J, Goadrich M. 2006. The relationship between Precision-Recall and ROC curves, p. 233–240. *In* Proceedings of the 23rd international conference on Machine learning.
193. Abu-Ali GS, Mehta RS, Lloyd-Price J, Mallick H, Branck T, Ivey KL, Drew DA, DuLong C, Rimm E, Izard J, others. 2018. Metatranscriptome of human faecal microbial communities in a cohort of adult men. Nat Microbiol 3:356–366.
194. Grant RF. 1998. Simulation of methanogenesis in the mathematical model ecosys. Soil Biol Biochem 30:883–896.
195. Riley WJ, Subin ZM, Lawrence DM, Swenson SC, Torn MS, Meng L, Mahowald NM, Hess P. 2011. Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM. Biogeosciences 8:1925.
196. Zhang G, Tian J, Jiang NA, Guo X, Wang Y, Dong X. 2008. Methanogen community in Zoige wetland of Tibetan plateau and phenotypic characterization of a dominant uncultured methanogen cluster ZC-I. Environ Microbiol 10:1850–1860.
197. Berestovskaya JJ, Kotsyurbenko OR, Tourova TP, Kolganova T V, Doronina N V, Golyshin PN, Vasilyeva L V. 2012. Methylorosula polaris gen. nov., sp. nov., an aerobic, facultatively methylotrophic psychrotolerant bacterium from tundra wetland soil. Int J Syst Evol Microbiol 62:638–646.
198. Dalcin Martins P, Hoyt DW, Bansal S, Mills CT, Tfaily M, Tangen BA, Finocchiaro RG, Johnston MD, McAdams BC, Solensky MJ, others. 2017. Abundant carbon substrates drive extremely high sulfate reduction rates and methane fluxes in prairie pothole wetlands. Glob Chang Biol 23:3107–3120.
199. Conrad R. 2020. Importance of hydrogenotrophic, aceticlastic and methylotrophic methanogenesis for methane production in terrestrial, aquatic and other anoxic environments: A mini review. Pedosphere 30:25–39.

200. Bridgham SD, Cadillo-Quiroz H, Keller JK, Zhuang Q. 2013. Methane emissions from wetlands: biogeochemical, microbial, and modeling perspectives from local to global scales. Glob Chang Biol 19:1325–1346.
201. Borrel G, Jézéquel D, Biderre-Petit C, Morel-Desrosiers N, Morel J-P, Peyret P, Fonty G, Lehours A-C. 2011. Production and consumption of methane in freshwater lake ecosystems. Res Microbiol 162:832–847.
202. Bastviken D, Tranvik LJ, Downing JA, Crill PM, Enrich-Prast A. 2011. Freshwater methane emissions offset the continental carbon sink. Science (80-) 331:50.

Appendices

Appendix A: Chapter 2 data tables including geochemistry, genome statistics, and abundance information.

See supplemental file "APPENDIXA_Methanohalophilus.xlsx" for Appendix A:

Tab 1: Genome relative abundance of *Methanohalophilus* across wells (location), time (days post hydraulic fracturing), chloride concentrations (grams per liter) and methylamine concentration.

Tab 2: Genome statistics table including genome completion, genome overages, GC content, length, number of proteins and number of core genes.

Tab 3: Anvi'o gene clusters denoting core and flexible gene clusters.

Tab 4: CRISPR-Cas system and array information for each isolate.

Tab 5: Extension of Figure 2.9 with all *Methanohalophilus* genomes.

Appendix B: Chapter 3 supplementary text including additional details on osmoprotectants, the Stickland reactions, and viruses.

Osmoprotection Activity

Hydraulic fracturing creates a unique environment for life in which salinities increase from freshwater to brine through time. In Utica well 1 sampled, salinities reached up to 95 g/L chloride (Appendix C). The pressure of salinity is reflected in the microcosm microbial community, as the genomes recovered are halotolerant and have the genomic potential to cope with elevated osmolarity (Figure 3.6). Further, both mechanisms for osmoprotection, including the salt-in strategy and production of compatible solutes, are detected in the metaproteome (1-2) (Figure 3.6).

Halanaerobium use the salt-in strategy, specifically utilizing multiple copies of sodium/proton antiporters (nhaC) that regulate intracellular sodium concentration while also balancing the number of protons in the cell (1). Contrary to prior reports (1), *Halanaerobium* in this microcosm also actively import and synthesize known osmoprotectants (Figure 3.6). We show that choline, proline, and glutamine are being actively imported or synthesized by *Halanaerobium* with no mechanism for degradation (Figure 3.6). While proline and glutamine could be assimilated by the cell, choline is imported and neither degradation mechanism (choline lyase or choline dehydrogenase) is present in the proteome or genome. Furthermore, no published *Halanaerobium* genomes, isolates or from metagenomics, have the genomic potential to degrade choline. *Halanaerobium* strains can also import maltose and trehalose via ABC transporters, but these compounds are actively degraded to D-glucose by maltose phosphorylase and alpha, alpha-trehalose phosphorylase, respectively. Similarly, *Halanaerobium* can transport and degrade glycine betaine, making it an unlikely osmoprotectant. It should be noted that taurine and mannitol are likely not being imported into

the cell because these compounds are not detected in the microcosm or in Utica produced fluids. Likewise, the respective transporters are not substrate specific and are able to import glycine betaine and fructose, respectively (Figure 3.6).

Like *Halanaerobium*, *Ca.* Uticabacter, employs both the salt-in and compatible solute strategy simultaneously. *Ca.* Uticabacter utilizes sodium/ proton antiporter (nhaC) and uptakes glycine betaine, proline, and glutamine, but does not degrade these compounds, suggesting that it is using them for osmoprotection (Figure 3.6). It is also possible that proline and glutamine are being used in protein synthesis. *Methanohalophilus* actively imports and synthesizes glycine betaine for osmoprotection from glycine and sarcosine by glycine and sarcosine methyltransferases, respectively (Figure 3.6). *Geotoga* uptakes trehalose, maltose, glutamine and glycine betaine, with glycine betaine and the sugars being the likely compatible solutes. Maltose and trehalose are being interconverted via maltose alpha-D-glucosyltransferase by *Geotoga* but the proteins for degradation are not detected. Notably, the only two osmoprotectants detected in the produced fluids from the Utica well time series were glycine betaine and choline, suggesting that these two amines are key in microbial salinity tolerance (Appendix C). Furthermore, based on glycine betaine trends, we infer both utilization (for osmoprotection, energy generation, and carbon and nitrogen assimilation) and production (for osmoprotection) of glycine betaine.

Viruses

Viruses accounted for 0.9% of the total microcosm metagenomic reads, denoting their prevalence in this *in vitro* ecosystem. Notably, viral peptides were detected in the metaproteomics data. Our microcosom proteomic data also provided evidence for the activity of both viral lifestyles. Evidence for virion-producing active infections, as opposed to a lysogenic state, was

provided by detection of multiple peptides for capsid production (e.g. terminase and head proteins). Also, we have evidence that some viral members are entering the lysogenic cycle and integrating themselves into host genome, as viral recombinase and resolvase proteins were also expressed.

Using nucleotide frequency (3), we demonstrated that viruses could be associated with every host. We predicted *Halanaerobium* was the most likely host for 8 viruses, *Methanohalophilus* for 2 viruses, *Ca.* Uticabacter for 4 viruses, and *Geotoga* for 2 viruses. Coordinated host and viral abundance patterns over time revealed no significant differences due to glycine betaine amendment, suggesting this treatment had little impact on viral predation. For three of the four microbial members, the microcosm viruses exhibited the same dynamics as their hosts over time (Figure 3.7). Alternatively, for *Methanohalophilus* and its most abundant associated virus, there was a clear decoupling between host and virus abundance patterns over time regardless of amendment.

To more directly link host and viral population genomes, we performed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) array analysis. Two of our microbial members had CRISPR arrays, with *Methanohalophilus* encoding 148 spacers in two CRISPR arrays and *Halanaerobium* encoding 206 spacers in four CRISPR arrays encoded a CRISPR-Cas system. None of the spacer sequences within the *Methanohalophilus* arrays matched viral genomes in our microcosms, suggesting that these spacers likely reflected historical viral encounters. We were able to link 8 *Halanaerobium* spacer sequences to 4 microcosm viral populations. Additionally, 14 of these Halanaerobium spacers also linked to 8 viruses recovered in the well used for this inoculum, as well as 2 viruses from another previously published well (4).

To predict if viral predation was ongoing in microcosms, we examined hosts for expression of CRISPR-Cas immunity genes. CRISPR-Cas proteins for the three functional stages of adaptive immunity were detected (adaptation, expression, interference) (Figure 3.8). Both *Methanohalophilus* and *Halanaerobium* expressed proteins for the adaptive stage of immunity (Cas1), suggesting active incorporation of spacers into the CRISPR loci. Concurrently, both also expressed proteins for the interference stage of viral immunity. *Halanaerobium* proteins were detected for the first part of the interference stage (Cas 5), which is implicated in producing cognate RNA that binds to invading DNA (5). Alternatively, *Methanohalophilus* proteins were detected for both parts of the interference stage (Cas 5 and Cas3), including cleavage of foreign viral DNA (5). Proteins for the expression stage (Cas6) were only detected from *Methanohalophilus* (Figure 3.8).

Stickland Reaction

The Stickland reaction is characterized by the oxidation of one amino acid coupled to the reduction of another (6,7). Since the discovery of this metabolism in 1934, several other non-amino acids have been characterized to take part in this reaction, including glycine betaine, sarcosine, and ornithine (7-9). Organisms use this metabolism to generate energy in the form of ATP via substrate level phosphorylation (10). Several organisms have been described to take part in this reaction, most of them members of the class Clostridia (9-10). Here we describe an active Stickland reaction in *Halanaerobium* and *Candidatus* Uticabacter.

Both *Halanaerobium* and *Candidatus* Uticabacter in the microcosm experiment use reductase mechanisms related to the glycine reductase mechanism (9). This family of reductase systems can reduce glycine, sarcosine, proline, or glycine betaine. In each system, there are generally three proteins: protein A (encoded by *grdA*), protein B (encoded by different genes based

on substrate specificity), and protein C (encoded by *grdCD*) (9). In the microcosm, *Halanaerobium* use the glycine betaine specific protein B (GrdHI*)*, while *Candidatus* Uticabacter use both the sarcosine and glycine specific protein B (GrdFG and GrdBE, respectively). Methods for determining reductase specificity were reported previously (4). Briefly, alignments of the GrdE/I/G/PrdA homolog amino acid sequences from our metagenomic database and known GrdE/I/G/PrdA from *Eubacterium acidominophilum* and *Clostridium sticklandii* revealed that the *Halanaerobium* homolog lacked a conserved cysteine residue and formed a monophyletic clade with other known glycine betaine specific reductases (Figure 3.10). Furthermore, two *Candidatus Uticabacter* homologs clustered with known sarcosine and glycine reductases. Both organisms actively employ these reductase mechanisms, with all proteins detected in the proteome. For *Halanaerobium*, a bin likely composed of multiple strains, there are two full mechanisms turned on: grdA (scaffold_194_1, scaffold_93_1), grdH (scaffold_194_3, scaffold_93_3), grdI (scaffold_194_2, scaffold_93_2), grdC (scaffold_69_9, scaffold_93_9), and grdD (scaffold_69_8, scaffold_93_10). *Ca.* Uticabacter used a sarcosine reductase and a glycine reductase: grdA (scaffold_169_8, scaffold_23_32), grdB (glycine specific, scaffold_169_4), grdE (glycine specific, scaffold_169_3), grdG (sarcosine specific, scaffold_23_26), grdF (sarcosine specific, scaffold_23_27), grdC (scaffold_169_9, scaffold_23_33), and grdD (scaffold_169_10, scaffold_23_34). For *Candidatus* Uticabacter, all proteins were detected except for GrdA. Given that GrdA was detected in low amounts relative to the rest of the operon for the highly abundant *Halanaerobium*, we posit that *Candidatus* Uticabacter is likely using the sarcosine reductase and GrdA is just below detection. *Candidatus* Uticabacter using multiple reductase mechanisms has been found previously in other organisms including *C. sticklandii* and *C. difficile* (9). Moreover, this finding is consistent with the only other published genome from this genus

(*Dethiosulfatibacter aminovorans* DSM 17477) has the genomic potential for three reductase mechanisms specific to glycine, sarcosine, and glycine betaine (Figure 3.10). One possible source of sarcosine is creatine through enzyme creatinease, which is expressed in *Ca.* Uticabacter.

Several reductants can be used to reduce glycine betaine, sarcosine and glycine. Here we show that *Halanaerobium* use lysine, serine, threonine, glycine, methionine, glutamate and alanine, while *Ca.* Uticabacter uses glutamate, leucine, phenylalanine, glycine and threonine. The oxidation of one amino acid in the Stickland reaction provides reducing power for the reduction of another amino acid. The key enzyme in the generation of this reducing power for each reductant follows: lysine (3,5-diaminohexanoate dehydrogenase, E.C. 1.4.1.11), serine (serine dehydratase, E.C. 4.3.1.17), threonine (threonine dehydratase, E.C. 4.3.1.19 and threonine dehydrogenase E.C. 1.1.1.103), glycine (glycine cleavage system), methionine (methionine gamma-lyase, E.C. 4.4.1.11), glutamate (glutamate dehydrogenase, E.C. 1.4.1.4), alanine (alanine dehydrogenase, E.C. 1.4.1.1), and leucine (leucine dehydrogenase, E.C. 1.4.1.9) (10).

These reductants could account for about 39% of glycine betaine reduced from T_0 to T_M, with lysine (17%), serine (7.2%), threonine (3.8%), glycine (4.1%), and methionine (6.7%) of glycine betaine reduction (Appendix C). Although glutamate and alanine are likely reductants in the Stickland reaction with glycine betaine, as the respective dehydrogenases were detected in the *Halanaerobium* proteome, these were not apparent by metabolite analyses, suggesting that alanine and glutamate are being synthesized more quickly than *Halanaerobium* is oxidizing them (Appendix C).

Lysine and glycine betaine are the most likely Stickland pair in the microcosm. Lysine is oxidized to acetate, butyrate and ammonia through crotonyl-CoA, with the key enzyme for the Stickland reaction being 3,5-diaminohexanoate dehydrogenase (10). This enzyme is active

concomitant with the glycine betaine reductase mechanism with the highest detection at T_M in the glycine betaine microcosm. Metabolites confirm the oxidation of lysine, as lysine is reduced by 93% overtime and accounts for 17.1% of glycine betaine reduction from T_0 to T_M (Figure 3.9). Moreover, butyrate is produced (8.13±0.5 μmoles) in a nearly 1 to 1 ratio with lysine loss (7.8±0.5 μmoles) from T_0 to T_F, congruent with lysine oxidation.

Glycine Cleavage System

As discussed previously, glycine is used as a Stickland oxidant (*Ca.* Uticabacter), a Stickland reductant (*Ca.s* Uticabacter and *Halanaerobium*), and also in osmoprotectant synthesis (*Methanohalophilus*) (Figure 3.11). This multi-enzyme complex oxidizes glycine to CO_2 and methylene-THF (11). Although the reaction can be ran in reverse, we hypothesize that *Halanaerobium* and *Ca.* Uticabacter are oxidizing the metabolite, freeing electrons to complete the Stickland reaction. Metabolites confirm this finding in the glycine betaine amended microcosm as 265.7±6.3 μmoles of glycine is depleted to 21.1±1.0 μmoles from T_0 to T_F. Moreover, we speculate that *Geotoga*, runs the glycine cleavage system in reverse, producing glycine because metabolites show glycine production from T_M to T_F in the no glycine betaine microcosm, when *Geotoga* activity is highest (Figure 3.1).

Ethanolamine Utilization

Halanaerobium employs a mechanism for ethanolamine utilization (Figure 3.11). Congruently, ethanolamine was detected in every time point of Utica produced fluids sampled here (Appendix C). In the microcosm, *Halanaerobium* converts ethanolamine, present in the produced fluid inoculum, into acetaldehyde and ammonium by using the ethanolamine ammonia lyase (EutBC, 4.3.1.7) (Figure 3.11, Figure 3.13). Acetaldehyde is then converted into acetyl-aldehyde by the aldehyde oxidoreductase (EutE) and subsequently to acetate through acetylphosphate.

Alternatively, acetaldehyde can be converted to ethanol by an alcohol dehydrogenase (EutG), which is often thought to be used as a detox mechanism (12) (Figure 3.13). Microcosm metabolites confirm this metabolism, as ethanolamine is reduced from a concentration of 165.3 ± 7.4 μmoles and 119.0 ± 27.4 μmoles to below detect in glycine betaine and no glycine betaine microcosms, respectively (Figure 3.11). In the both the glycine betaine and no glycine betaine enrichment, EutE is detected at higher levels than EutG, suggesting that *Halanaerobium* is using ethanolamine for energy, rather as a detoxification mechanism.

Halanaerobium-encoded detected proteins for ethanolamine utilization include: ethanolamine ammonia lyase large subunit (EutB, scaffold_31_26), ethanolamine ammonia lyase small subunit (EutC, scaffold_31_25), acetylaldehyde dehydrogenase (EutE, scaffold_31_22), alcohol dehydrogenase (EutG, scaffold_31_10), microcompartments/ carboxysome structural proteins (scaffold_31_14, scaffold_31_21, scaffold_31_24), ethanolamine transporter (EutH, scaffold_31_13). All proteins detected in *Halanaerobium* proteome for the Eut operon are shown in Figure 3.13. Ethanolamine ammonia lyase is a vitamin B12 requiring enzyme, thus *Halanaerobium* imports this cofactor via transporters and does not make it *de novo*. We note that ethanolamine transporter protein EutH is detected in low levels and that ethanolamine is likely diffusing across the membrane concurrent with transport (12) (Figure 3.13).

Methane and Acetate Mass Balance Calculations

Given the importance of glycine betaine to hydraulically fractured shale organisms, both as a substrate and an osmoprotectant, and the presence of glycine betaine in the Utica well sampled, we amended produced fluids with glycine betaine and tracked microbial activity and metabolites through time (Figure 3). In the microcosm, *Halanaerobium* utilized glycine betaine reductase to reduce glycine betaine to TMA (*grdHI*), which was most active at T_M in the glycine

betaine amended microcosm. Analysis of metabolites by NMR support the proteomics data, showing that in the glycine betaine amended microcosm 42.1±2.4 μmoles of glycine betaine was reduced to 37.9±0.6 μmoles of TMA from T_0 to T_M, a 90% reduction (Figure 3). Similarly, in the no glycine betaine microcosm, 2.6±0.1 μmoles of glycine betaine was 81.2% reduced to TMA from T_0 to T_M (Figure 3.9).

The TMA produced by *Halanaerobium* is utilized by *Methanohalophilus*, a methylotrophic methanogen (Figure 3.9). The most methane is produced from T_M to T_F in the glycine betaine amended microcosm (Figure 3.9). From T_M to T_F, 95% and 60% of TMA is converted to methane in the glycine betaine and no glycine betaine microcosm, respectively (Figure 3.9). Congruently, the most *Methanohalophilus* proteins are detected in T_F timepoints, with the glycine betaine amended microcosm having statistically more than the no glycine betaine microcosm. Furthermore, the trimethylamine specific pyrrolysine-containing methyltransferase (MttB) and the corresponding corrinoid protein (MttC) were highly detected in T_F in the glycine betaine amended microcosm, statistically more than in any other sample. Methyltransferase proteins specific to dimethylamine, monomethylamine, and methanol and all proteins necessary for methanogenesis were also detected (Appendix C). Dimethylamine and monomethylamine concentrations followed the same pattern as trimethylamine, increasing from T_0 to T_M and decreasing in from T_M to T_F (Appendix C). If we assume all methane production was fueled indirectly by glycine betaine, 72% of glycine betaine accounts for all methane produced in the glycine betaine amended microcosm from T_0 to T_F (Figure 3.9). There was no potential for glycine betaine or choline demethylation in our microcosm experiments, as no non-pyrrolysine trimethylamine methyltranserfases were detected (13-14).

Acetate, also produced one to one with TMA in the reduction of glycine betaine, had a net increase of 63.8±1.5 μmoles from T_0 to T_M in the glycine betaine amended microcosm (Figure 3.11). The excess acetate (25.9±1.5 μmoles) produced in the glycine betaine amended microcosm can be accounted for by residual carbon fermentation, as the no glycine betaine microcosm produced 24.0±1.7 μmoles acetate, of which only 2.5±0.1 μmoles came from glycine betaine fermentation. Given that acetate is produced in a one to one stoichiometric balance with TMA from glycine betaine reduction (9), we know that excess acetate (not accounted for by glycine betaine reduction, 25.9±1.5 μmoles) was produced in the glycine betaine amended microcosm. Notably this accounts for ~97% of acetate in non-amended microcosm (24.0±1.7 μmoles), where no glycine betaine was added. With glycine betaine accounting for 46% of acetate production, the excess can be accounted for through sugar fermentation, with glucose (2.3%), trehalose (21.1%), ethylene glycol (11.1%), ethanolamine (1.4%), pyruvate (0.3%), maltose (2.3%), and fructose (4.0%) accounting a substantial portion of acetate production in the amended glycine betaine microcosm. See Appendix C for detailed acetate mass balance calculations.

Back to the field: Validation of microcosm generated hypotheses across wells

We compared our metabolic findings from microcosm experiments to previously published hydraulically fractured shale datasets and 33 metagenomes paired to metabolites published here. Prior to the Daly, *et al.* study, HF microbiology studies were limited to 16S rRNA analyses, did not have time series data including injected fluids, or did not include metabolites (15-17). Given that Daly *et al.* was a single well, it was necessary to apply our microcosm findings to other wells in different shale formations. Here we add 33 metagenomes and paired metabolites to build a HF database of 38 metagenomes. The 33 additional metagenomes came from injected fluids and

produced fluids from four wells in the Marcellus and Utica shales. Two Utica wells were located in Ohio, two Marcellus wells in West Virginia, and one Marcellus well in Pennsylvania (Figure 3.14). Chloride concentrations increased over time in all wells (Figure 3.14). Metabolites and metagenome information can be found in Appendix C.

In light of the importance of the Stickland reaction to hydraulically fractured shale organisms, we mined the published isolate genomes and metagenomes from produced fluids for the necessary genes (4,18-19). We found that 24% of genomes in our shale database had the potential to use glycine betaine, 5 of them *Halanaerobium*. Moreover, we found that the most abundant *Halanaerobium* strain at late time points in the well sampled here has a GrdI (*Halanaerobium* 6-U2, genome previously published in Booker, et al. (18). As previously reported, *Frackibacter*, a new genus within the Halobacteroidaceae discovered in shale, has the potential to reduce glycine betaine (4), and 2 of 3 publicly available *Frackibacter* genomes have the genomic potential to use glycine betaine.

Appendix C References

1. Oren A (1999) Bioenergetic aspects of halophilism. *Microbiology and molecular biology reviews* 63(2):334-348.
2. Sleator RD & Hill C (2002) Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS microbiology reviews* 26(1):49-71.
3. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, & Sun F (2016) Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic acids research* 45(1):39-53.
4. Daly RA*, et al.* (2016) Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nature microbiology* 1:16146.

5. Makarova KS, *et al.* (2011) Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology* 9(6):467-477.

6. Stickland LH (1934) Studies in the metabolism of the strict anaerobes (genus Clostridium): The chemical reactions by which Cl. sporogenes obtains its energy. *Biochemical Journal* 28(5):1746.

7. Nisman B (1954) The stickland reaction. *Bacteriological reviews* 18(1):16.

8. Naumann E, Hippe H, & Gottschalk G (1983) Betaine: new oxidant in the Stickland reaction and methanogenesis from betaine and L-alanine by a Clostridium sporogenes-Methanosarcina barkeri coculture. *Applied and environmental microbiology* 45(2):474-483.

9. Andreesen JR (2004) Glycine reductase mechanism. *Current opinion in chemical biology* 8(5):454-461.

10. Fonknechten N, *et al.* (2010) Clostridium sticklandii, a specialist in amino acid degradation: revisiting its metabolism through its genome sequence. *Bmc Genomics* 11(1):555.

11. Andreesen JR (1994) Glycine metabolism in anaerobes. *Antonie Van Leeuwenhoek* 66(1-3):223-237.

12. Garsin DA (2010) Ethanolamine utilization in bacterial pathogens: roles and regulation. *Nature Reviews Microbiology* 8(4):290-295.

13. Ticak T, Kountz DJ, Girosky KE, Krzycki JA, & Ferguson DJ (2014) A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. *Proceedings of the National Academy of Sciences* 111(43):E4668-E4676.

14. Watkins AJ, Roussel EG, Webster G, Parkes RJ, & Sass H (2012) Choline and N, N-dimethylethanolamine as direct substrates for methanogens. *Applied and environmental microbiology* 78(23):8298-8303.

15. Mouser PJ, Borton M, Darrah TH, Hartsock A, & Wrighton KC (2016) Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. *FEMS microbiology ecology* 92(11).

16. Cluff MA, Hartsock A, MacRae JD, Carter K, & Mouser PJ (2014) Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus Shale gas wells. *Environmental science & technology* 48(11):6508-6517.

17. Akob DM, Cozzarelli IM, Dunlap DS, Rowan EL, & Lorah MM (2015) Organic and inorganic composition and microbiology of produced waters from Pennsylvania shale gas wells. *Applied Geochemistry* 60:116-125.

18. Booker AE*, et al.* (2017) Sulfide Generation by Dominant Halanaerobium Microorganisms in Hydraulically Fractured Shales. *mSphere* 2(4):e00257-00217.

19. Lipus D, Vikram A, Ross DE, & Bibby K (2016) Draft genome sequence of Methanohalophilus mahii strain DAL1 reconstructed from a hydraulic fracturing-produced water metagenome. *Genome announcements* 4(5):e00899-00816.

Appendix C: Chapter 3 data tables including geochemistry, genome statistics, mass balance calculations, and key genes.

See supplemental file "APPENDIXC_GBData.xlsx" for Appendix C.

Tab 1: Detected metabolites (μM) and chloride (mg/L) from field time series (n=41) collection. Concentrations of zero denote that the metabolite was below detection.

Tab 2: Detected metabolites (μM) from microcosm experiments (n=21). Concentrations of zero denote that the metabolite was below detection.

Tab 3: Microcosm mass balance calculations.

Tab 4: Optical density and gas chromatography measurements through time in the microcosm experiment.

Tab 5: Table of metagenome stats for produced fluids and microcosm sequencing.

Tab 6: Metagenome assembled bacterial and archaeal genome quality statistics.

Tab 7: Metagenome assembled viral genome quality statistics.

Tab 8: Scaffold and gene information for key metabolisms discussed.

Tab 9: NSAF values for each protein detected in the microcosms by metaproteomics.

Tab 10: Strain resolved microbial abundances (by ribosomal S3 protein) across input and produced fluid samples.

Tab 11: Value Importance in Projection for each predicted metabolite in Figure 3.14.

Appendix D: Chapter 4 data tables including cohort statistics, metabolite concentrations, genome statistics, GAMAdb entries, and genome DRAM annotations.

See supplemental file "APPENDIXD_GutData.xlsx" for Appendix D.

Tab 1: Cohort statistics including sex, age, weight, BMI, smoking status, meat consumption, etc.

Tab 2: Urine and fecal metabolites for patient cohort.

Tab 3: 16S rRNA taxonomic abundance for patient cohort.

Tab 4: Metagenome sequencing information including size (Gbp), reads, and accession numbers.

Tab 5: Relative abundance of MAGs recovered from human gut metagenome samples in this study, utilized in Figure 4.11E.

Tab 6: Quality and taxonomy information recovered from human gut metagenome samples in this study.

Tab 7: Important genes for ACVD predictions.

Tab 8: GAMAdb entries, including genome, taxonomy, and gene type.

Tab 9: DRAM summary annotation file for genomes in GAMAdb.

Tab 10: Microcosm metaproteomic data.

Tab 11: Microcosm metabolite data.