Title

Microbial Tapestry: Interwoven activities of novel uncultivated microorganisms, their genome structures, and mobile genetic elements in subsurface ecosystems

Permalink

https://escholarship.org/uc/item/1h03z5cv

Author

Valentin-Alvarado, Luis Enrique

Publication Date

2024

Peer reviewed|Thesis/dissertation

Microbial Tapestry: Interwoven activities of novel uncultivated microorganisms, their genome structures, and mobile genetic elements in subsurface ecosystems

by

Luis Valentin-Alvarado

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jill F. Banfield, Co-chair
Professor David F Savage, Co-chair
Professor Michiko Taga
Professor Jennifer Doudna

Spring 2024

Abstract

Microbial Tapestry: Interwoven activities of novel uncultivated microorganisms, their genome structures, and mobile genetic elements in subsurface ecosystems

by

Luis Valentin-Alvarado

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Co-chair
Professor David Savage, Co-chair

Recent advances in metagenomics have made it possible to resolve genomes of individual microorganisms from metagenomic data. This is known as genome-resolved metagenomics (GRM), and it has provided new insights into the ecological roles of individual microbial species within complex communities. GRM has enabled the discovery of novel microbial groups, including the Candidate Phyla Radiation (CPR) bacteria and the Asgard archaea. CPR bacteria are widespread and often live in symbiotic relationships with other microorganisms, while Asgard archaea are thought to be related to the ancestor of eukaryotic cells. My dissertation focuses on these newly resolved microbial groups, including consideration of their lifestyles, interactions, and evolution. Specific focus includes the close and specific inter-organism interactions (such as CPR bacteria and their host microorganisms) and evolutionary processes in archaea that may have laid the foundations for the development of multicellularity within the Asgard archaea. A central aspect of this research is the consideration of Mobile Genetic Elements (MGEs) that associate with these microbial groups and the biochemical mechanisms that regulate interactions between MGEs and their microbial hosts.

The first chapter of my dissertation focuses on Candidate Phyla Radiation (CPR) bacteria, which are commonly found in microbial communities but their role in biogeochemical cycles is not well understood. I studied biofilms that grow in sulfide-rich springs and found that these host diverse CPR bacteria. Ultra-small cells attached to the surfaces of chemolithotrophic sulfur-oxidizing filamentous bacteria are inferred to be CPR bacterial episymbionts (the first reported association of CPR with Proteobacteria). Some CPR bacteria have a novel electron bifurcating group 3b [NiFe]-hydrogenase, and other proteins potentially linked to sulfur and hydrogen metabolism. Thus, these CPR have the potential to directly impact biogeochemistry in groundwater systems. The second chapter focuses on hydrogenases—enzymes crucial for hydrogen cycling in anaerobic ecosystems. Through a combination of metagenomics, biochemistry, and genomics, this work reveals previously unknown [FeFe]-hydrogenases in novel archaeal lineages. This discovery not only expands our understanding of these vital enzymes but also highlights the metabolic versatility and evolutionary innovation within uncultivable archaeal phyla.

The third and fourth chapters focus on Asgard archaeal metabolism and mobile genetic elements. Through metagenomic analysis of wetland soil samples, I report two complete genomes for Atabeyarchaeota (after Atabey, a supreme goddess of freshwater and fertility in the Taíno religion from Puerto Rico), a new Asgard archaeal lineage, and the first complete genome of a Freyarchaeota, providing insights into their metabolic potential, evolutionary history, and ecological roles. I also investigated their diverse mobile genetic elements, which contribute to microbial adaptation and evolution. This is the first focused study of Asgard MGEs and thus of the processes that drive genome divergence and population dynamics in the modern lineages. The fifth chapter discusses the challenges and opportunities in archaeal genomics, emphasizing the limitations of our current knowledge due to the scarcity of cultivated strains and complete genomes. With the advent of long-read sequencing technologies, such as PacBio and Nanopore, I describe my efforts in sequencing wetland soil samples, which led to the recovery of hundreds of circular genomes. This endeavor not only highlights the diversity and metabolic potential of archaea but also underscores the importance of advanced sequencing technologies in uncovering the hidden diversity of microbial life.

In summary, these studies provide new insights into the diversity and evolution of CPR bacteria and Asgard archaea, as well as the role of mobile genetic elements in shaping microbial genomes. The application of genome-resolved metagenomics has enabled a more comprehensive understanding of microbial communities and their functions in complex environments, which has important implications for fields such as microbial ecology, evolution, biotechnology, and population dynamics.

Spanish translation

Los recientes avances en la metagenómica han permitido resolver los genomas de microorganismos individuales a partir de datos metagenómicos. Esto se conoce como metagenómica de resolución genómica (GRM), y ha proporcionado nuevos conocimientos sobre los roles ecológicos de especies microbianas dentro de comunidades complejas. La GRM ha permitido el descubrimiento de nuevos grupos microbianos, incluyendo las bacterias de la Radiación de Filos Candidatos (CPR por sus siglas en inglés: Candidate Phyla Radiation) y las arqueas Asgard. Las bacterias CPR están muy extendidas y a menudo viven en relaciones simbióticas con otros microorganismos, mientras que se cree que las arqueas Asgard están relacionadas con el ancestro que dio lugar a las células eucariotas. Mi tesis doctoral se centra en estos grupos microbianos recientemente resueltos, incluyendo sus estilos de vida, interacciones y evolución. Los enfoques específicos incluyen las interacciones inter-organismos cercanas y específicas (como las bacterias CPR y sus microorganismos hospedadores) y los procesos evolutivos en las arqueas que pudieron haber sentado las bases para el desarrollo de la multicelularidad. Asgard. Un aspecto central de esta investigación es la consideración de los Elementos Genéticos Móviles (MGEs por sus siglas en inglés: Mobile Genetic Elements) que se asocian con estos grupos microbianos y los mecanismos bioquímicos que regulan las interacciones entre los MGEs y sus hospedadores microbianos.

El primer capítulo de ésta tesis doctoral se centra en las bacterias CPR, que se encuentran comúnmente en las comunidades microbianas, pero actualmente no queda claro cuál es su

contribución en los ciclos biogeoquímicos. Estudié biopelículas que crecen en manantiales ricos en sulfuro y encontré que estas albergan diversas bacterias CPR. Las células ultra pequeñas adheridas a las superficies de bacterias filamentosas quimiolitotróficas oxidantes de azufre se infieren como episimbiontes bacterianos CPR (la primera asociación reportada de CPR con Proteobacterias). Algunas bacterias CPR tienen una nueva hidrogenase del grupo 3b [NiFe] bifurcadora de electrones, y otras proteínas potencialmente vinculadas al metabolismo del azufre y el hidrógeno. Por lo tanto, estas CPR tienen el potencial de tener un impacto directo en la biogeoquímica en los sistemas de aguas subterráneas. El segundo capítulo se enfoca en las hidrogenasas, enzimas cruciales para el ciclo del hidrógeno en ecosistemas anaeróbicos. A través de una combinación de metagenómica, bioquímica y genómica, este trabajo revela hidrogenasas [FeFe] previamente desconocidas en nuevos linajes de arqueas. Este descubrimiento no solo amplía nuestra comprensión de estas enzimas vitales, sino que también resalta la versatilidad metabólica y la innovación evolutiva dentro de los filos arquéales no cultivables.
El tercer y cuarto capítulos se enfocan en el metabolismo de las arqueas Asgard y los MGEs. A través del análisis metagenómico de muestras de suelo de humedales, en esta tesis se reportan dos genomas completos para Atabeyarchaeota (después de Atabey, una diosa suprema de las aguas dulces y la fertilidad en la religión Taína de Puerto Rico), un nuevo linaje de arqueas Asgard, y el primer genoma completo de un Freyarchaeota, proporcionando información sobre su potencial metabólico, historia evolutiva y roles ecológicos. En el cuarto capítulo investigué sus diversos MGEs, que contribuyen a la adaptación y evolución microbiana. Este es el primer estudio enfocado en los MGEs de Asgard y, por lo tanto, de los procesos que impulsan la divergencia genómica y la dinámica poblacional en los linajes modernos. El quinto capítulo discuto los desafíos y oportunidades en la genómica arqueal, enfatizando las limitaciones de nuestro conocimiento actual debido a la escasez de cepas cultivadas y genomas completos. Con las nuevas tecnologías de secuenciación de lectura larga, como PacBio y Nanopore, describo la secuenciación de muestras de suelo de humedales, lo que condujo a la recuperación de cientos de genomas circulares. Este capítulo no sólo resalta la diversidad y el potencial metabólico de las arqueas, sino que también resalta la importancia de las tecnologías de secuenciación de lecturas largas para descubrir la diversidad oculta de la vida microbiana.

En resumen, mi tesis doctoral nuevos conocimientos sobre la diversidad y evolución de las bacterias CPR y las arqueas Asgard, así como el papel de los MGEs en la configuración de los genomas microbianos. La aplicación de la metagenómica de resolución genómica ha permitido una comprensión más completa de las comunidades microbianas y sus funciones en entornos complejos, lo que tiene importantes implicaciones en campos como la ecología microbiana, la evolución, la biotecnología y la dinámica poblacional.

Dedicated to my mother,

I dedicate this dissertation to you with heartfelt gratitude.

To my mom, Lesly Alvarado Rivera, your unwavering love and support have been my rock, and this work is a tribute to your profound influence on my journey to this achievement.

Dedicado a mi madre,

Te dedico esta tesis con profunda gratitud.

Para mi mamá, Lesly Alvarado Rivera, tu amor y apoyo inquebrantables han sido mi roca, y este trabajo es un tributo a tu profunda influencia en mi camino hacia este logro.

Table of Contents

Introduction

It has been 20 years since the first reconstruction of complete genomes from metagenomes. This groundbreaking advancement has enabled us to explore the roles of microbes from a community-based perspective, illuminating novel microbial taxa with profound implications for biogeochemical cycles, climate change, and the origins of eukaryotes. This thesis explores the diversity and role of enigmatic microbes in subsurface environments, the nature of their mobile genetic elements, and uncovers novel enzymes relevant in Earth's biogeochemical cycles.

As a Ph.D. candidate at the University of California, Berkeley, in Dr. Jill Banfield's lab, my dissertation is a deep dive into the world of microbial communities, with a particular focus on the analysis of complete genomes from metagenomes. This research targeted chemoautotrophic microbial communities growing in discharging groundwater and in a wetland soil ecosystem. The roles of symbiotic organisms in communities, novel mechanisms of hydrogen cycling, anaerobic carbon compound metabolism and the diversity and dynamics of Mobile Genetic Elements (MGEs) and their hosts were specific foci.

Candidate Phyla Radiation (CPR) bacteria, a large group of ultra small celled organisms that likely adopt episimbiotic lifestyles, have garnered scientific interest due to their ubiquity in diverse ecosystems. However, their specific roles in biogeochemical cycles remain unclear. Chapter 1 examines the metabolic capacities and interdependencies within chemoautotrophically-based microbial biofilm communities, with emphasis on CPR bacteria. Utilizing genome-resolved metagenomics, geochemistry, X-ray spectromicroscopy, and scanning electron microscopy, it was found that these tiny bacterial cells associate with the surfaces of sulfur-oxidizing bacterial hosts that underpin primary production in discharging cold, sulfide-rich water. The reconstructed genomes revealed the potential of CPR to contribute to sulfur and hydrogen cycles. This research was published as a first author paper in the journal *Microbiome*.

Hydrogen is one of the most important energy resources in microbial ecosystems, yet little is known about the complexes that make and use hydrogen in archaea, especially those from lineages lacking isolated representatives. Chapter 2 presents a collaborative effort to describe archaeal hydrogenase diversity and structures. Through the integration of genome-resolved metagenomics and biochemistry, we elucidated previously unknown hydrogenases that are fusions of [FeFe]- and [NiFe]-hydrogenases in archaeal genomes. The research also discovered an Fe-Fe- ultra-minimal hydrogenase in DPANN archaea. I performed a search for novel hydrogenases in over 12,000 archaeal genomes, used state-of-the-art phylogenetic approaches and structural analysis to resolve the ancestral position of these novel hydrogenases, and performed metabolic reconstructions to elucidate the potential mechanisms of these enzymes in archaea from anoxic environments. This work was published as a co-first author paper in the journal *Cell*.

Asgard archaea are of great interest from the perspective of evolution as they have been identified as the probable ancestors of eukaryotes, the domain of life that includes plants, animals, and humans. Chapter 3 addresses the challenge of obtaining confident insights regarding the metabolic capacities of Asgard archaea through reconstruction and analysis of complete genomes. To augment the existing set of Asgard archaeal genomes from marine settings, the first complete genomes were recovered from anaerobic wetland soil and used to define one new phylum-level group that was named Atabeyarchaeota. The first complete genome from a sibling lineage, Freyarchaeota, was reconstructed and analyzed in the same study. Metatranscriptomic datasets were obtained to identify the *in situ* activities of Atabeyarchaeota and Freyarchaeota. [NiFe]-hydrogenases, pyruvate oxidation, and carbon fixation via the Wood-Ljungdahl pathways were expressed *in situ*, as were enzymes for amino acid metabolism, anaerobic aldehyde oxidation, hydrogen peroxide detoxification and glycerol and carbohydrate breakdown to acetate and formate. Overall, soil-associated Asgard archaea are predicted to be non-methanogenic, acetogens, likely impacting reservoirs of substrates for methane production in terrestrial ecosystems. The article of this work is a first author paper and is currently under review.

There are gaps in our understanding of the modes by which Asgard archaea evolve. Chapter 4 investigates the genetic repertoire of MGEs in Asgard archaea. Population genomic information for Atabeyarchaeota were used to identify the subset of MGE that are able to integrate and excise from the host chromosomes, providing confident host linkages for these MGE and their coexisting circularized variants. MGEs were classified based on inferred functions of their encoded proteins, in some cases making use of predicted *in silico* structures to identify protein functions that otherwise were impossible to decipher based on their primary sequences. Additionally, this chapter reports on genomically-encoded defense systems and methylation patterns that differentiate these archaea, enhancing our understanding of their evolution.

Chapter 5 addresses the current limitations in our knowledge of some major groups of archaea that are largely due to the small number of cultivars and the low quality of many draft genomes. High fidelity (HiFi) long-read PacBio sequencing was leveraged to genomically define microorganisms populating wetland soil samples. 1,200 very high quality genomes were reconstructed for bacteria, archaea and associated MGEs. Of these, over 500 genomes are circularized and were classified as either complete (when validated using Illumina reads) or potentially complete (complete pending minor error correction, such as single nucleotide polymorphisms). Our analyses focused on archaeal genomes from groups lacking very high quality genomes, such as Asgardarchaeota, Brockarchaeota, Bathyarchaeia, Thaumarchaeota, Hadearchaeales, Methanomethyliales, methanogens, and putatively symbiotic DPANN archaea. Phylogenetic analysis revealed that many of these genomes belong to new classes and are among the first representatives genomically described from wetland soils (e.g., Methanomethylicia and Brockarchaeota). The availability of complete genomes has enabled precise measurements of genome size, architecture, and gene content, as well as the accurate determination of linkages between genes and RNA expression. Metatranscriptomic datasets from the same samples were used to identify highly expressed genes, providing insights into the roles of these archaea in deep, wetland soil ecosystems. In combination with metabolic reconstruction, we infer that these archaea contribute to wetland soil carbon cycling via degradation of amino acids and complex carbohydrates, methanogenesis, carbon fixation and consumption, and production of hydrogen through a suite of diverse hydrogenases. The genomes will serve as important references for future research.

In conclusion, this dissertation advances our understanding of several novel microbial lineages and their likely roles in terrestrial processes. By combining analyses of complete and draft genomes from metagenomes with bioinformatics methods -that involve structure prediction to assign

function to highly divergent proteins- as well as transcriptomic activity measurements, the research sheds light on the metabolic capacities and functions of uncultivable archaea and CPR bacteria in the context of their communities. We unveiled a diverse repertoire of MGEs for groups of organisms for which few or none were known, expanding appreciation of the ways that these elements enable rapid shifts in genetic potential within populations and possibly contribute to longer term lineage evolution. In the future, some of these elements could be adapted for delivery of genome editing tools to genetically manipulate uncultivated organisms within the context of their communities. We also reported the genome methylation patterns in Asgard archaea and, in one case, link a methylation pattern to a methylase encoded in an integrated plasmid. Differences in methylation patterns and inventories of defense systems build out our appreciation of the ways in which mobile elements and their host organisms interactions unfold. Overall, the findings have implications for understanding of microbial processes that impact carbon storage and greenhouse gas emissions and the metabolism and modes of evolution of archaeal groups that share a common ancestor with eukaryotes.

Acknowledgments

I am deeply grateful to my mentor, Jill Banfield, for her constant support, thoughtful critiques, and patient guidance throughout my PhD journey. Her expertise and dedication have been crucial to the completion of this dissertation. Jill, your unwavering belief in me during challenging times has left a lasting impact, and I will always be thankful for your mentorship. Learning to prioritize data analysis from you has been a transformative experience, and your ability to navigate complex tasks with ease continues to inspire me. Your passion for science and relentless pursuit of illuminating the unknown have been the driving forces behind my decision to join your lab. The fieldwork experiences and mentorship you provided during those visits have significantly shaped my scientific journey.

I would also like to express my sincere gratitude to my co-mentor, Dave Savage, whose insightful feedback has pushed me to refine my thinking and elevate my work. His encouragement and rigorous standards have been instrumental in shaping this dissertation. Dave allowed me to further explore my potential as a bench scientist and helped me navigate the fascinating world of biochemistry in his lab. I have learned about a wide range of topics, from RubisCO and CRISPR to plants, from many members of his lab. I will always remember Dave's words during a walk through Berkeley when I asked, "When is it enough for a PhD?" His response, emphasizing the importance of acquiring the desired skills and the opportunity to continue learning during a postdoc, has stayed with me.

I extend my thanks to Michi Taga and Jennifer Doudna for being part of my doctoral dissertation and helping me from the beginning of my PhD until the end. I am incredibly grateful to my labmates from both the Banfield and Savage labs. Special thanks to Banfield lab members, especially Rohan Sachdeva, Lily Law, Jordan Hoff, Leylen Miloslavich, Raphaël Méheust, Daniel Gittins, Cindy Castelle, Christine He, Andreja Kust, Spencer Dimond, Lin-Xing Chen, Kaden DiMarco, Tom Hessler, Bethany Kolody, Jackie Zorz, Elliot Weiss, Kaitlin Creamer, Ling-dong Shi, Veronika Kivenson, Jack Kim, Saoirse Disney-McKeethen, Marie Schölmerich, and Jacob West-Roberts, for his guidance in bioinformatics and the countless hours spent teaching me programming. Jacob, you are a good friend and I feel so grateful that our paths crossed in grad school. I am thankful to Alexander Jaffe for being both a mentor and friend during my time as a PhD student. His guidance and support were invaluable in helping me navigate the challenges of graduate school and complete my dissertation successfully. Equally important are members from the Savage lab: Rob Nichols, Noam Prywes, Avi Flamholz, Jack Desmarais, Julia Borden, David Ding, Evan Grover, Brittney Thornton, Rachel Weissman, Christian Nixon, Maria Lukarska, Luke Oltroge, Christian Nixon, Jorge Rodriguez, Joseph Rivera, Naiya Phillips and Andrew Plebanek. Special thanks to Luke for teaching me about protein folding and for the engaging discussions on microbial life and novel metabolism.

I am indebted to my colleagues and the academic staff at the Plant and Microbial Biology Department for their invaluable input and assistance throughout this process. Their contributions have helped shape this dissertation, and their encouragement has kept me motivated. Special thanks to Rocio Sanchez and Lyn Rivera for their support. I am also grateful to my PMB cohort, especially Lorenzo Washington and Nicholas Karavolias. My friend Matt Metzger, with whom I embarked on the Ph.D. journey by driving from Boston to Berkeley, has become my best friend, travel partner, and a brother. I am grateful for his companionship and support. My most sincere thanks to Megan Hochstrasser, Hope Henderson, Andy Murdock, Clarice de Azevedo Souza, and Melanie Leavitt from the Innovative Genomics Institute for giving me the opportunity to be part of the DEI committee.

I appreciate the collaboration of Brett Baker, Michael Manga, Sirine Fakra, Alex Crits-Christoph, and Rohan Sachdeva. Rich Roberts, Daniel M. Portik, Jeremy E. Wilkinson, Siyuan Zhang and Pok Leung. Valeria De Anda, thank you for all your support during the last chapters of this dissertation, your expertise was very helpful. Thank you for the hours via Zoom reading the manuscripts together and obviously to your cat Sagan for its aggressiveness to end the meetings for a break. Gracias amiga!

I am grateful for the opportunity to have mentored two talented undergraduate students, Michael Cui and Astrid Quile. They taught me the value of mentorship and always went the extra mile, even in the most challenging moments.

I would like to acknowledge my scientific mentors: Chris Greening, for inspiring me to be the best version of myself and to advocate for mental health; Michi Taga and Arash Komeili, for believing in me from the very beginning; Dianne Newman, for her guidance in my scientific career. Since meeting her for the first time in Woods Hole in 2017, she has introduced me to many great scientists and helped pave my way to a PhD program. I am also grateful to my former mentors, Lilliam Casillas Martinez, Jorge Escalante-Semerena, Dan Repeta, Mak Saito, and Otto Cordero, Emily Balskus for providing me with scientific opportunities and helping me find my path in science. Special thanks to my high school teacher Maria E. Thiele Solivan for providing me with opportunities during the science fair projects.

On a personal note, I would like to express my deepest gratitude to my mother, Lesly Alvarado Rivera. As a single mother, you raised four children with unwavering dedication and love. This dissertation will never be enough to convey my appreciation for your countless sacrifices and the strength you have shown. To my titi Leyda Rivera Flores, thank you for everything you have done for me. I hope that I have made you proud. Your love, encouragement, and sacrifices have been instrumental in shaping me into the person I am today. To my grandparents, Maria Rivera Flores and Jose A. Pi, thank you for your love and support throughout my life. To Valdo, Javier, Antonio, Javier, Kevin, Franche, Suheil, Valdo, Luisito and Leydita thank you for being my pillars of strength and for always being there for me. Last but not least, I would like to express my heartfelt gratitude to my aunt, Sylvette Alvarado, for your unconditional love and support, which have been instrumental in my journey. Los amo y espero que estén muy orgullosos de mi.

"If I have seen further it is by standing on the shoulders of giants." - Isaac Newton

1. Autotrophic biofilms sustained by deeply-sourced groundwater host diverse bacteria implicated in sulfur and hydrogen metabolism

The following chapter is a modified version with permission of authors of the following published article: Valentin-Alvarado LE, Fakra S, Probst A, Giska JR, Jaffe A, Oltrogge LM, West-Roberts J, Rowland J, Manga M, Savage DF, Greening C, Baker BJ and Banfield JF Autotrophic biofilms sustained by deeply-sourced groundwater host diverse bacteria implicated in sulfur and hydrogen metabolism. *Microbiome* 12, 15 (2024).

Abstract

Biofilms in mineral sulfide-rich springs present intricate microbial communities that play pivotal roles in biogeochemical cycling. We studied chemoautotrophically-based biofilms that host diverse CPR bacteria and grow in sulfide-rich springs to investigate microbial controls on biogeochemical cycling. Sulfide springs biofilms were investigated using bulk geochemical analysis, genome-resolved metagenomics and scanning transmission x-ray microscopy (STXM) at room temperature and 87 Kelvin. Chemolithotrophic sulfur-oxidizing bacteria, including Thiothrix and or Beggiatoa, dominate the biofilms, which also contain CPR-affiliated Gracilibacteria, Absconditabacteria, Saccharibacteria, Peregrinibacteria, Berkelbacteria, Microgenomates, and Parcubacteria. STXM imaging revealed ultra-small cells near the surfaces of filamentous bacteria that may be CPR bacterial episymbionts. STXM and NEXAFS spectroscopy at carbon K and sulfur L2,3 edges show that filamentous bacteria contain protein-encapsulated spherical elemental sulfur granules, indicating that they are sulfur-oxidizers, likely Thiothrix. Berkelbacteria and Moranbacteria in the same biofilm sample are predicted to have a novel electron bifurcating group 3b [NiFe]-hydrogenase, putatively a sulfhydrogenase, potentially linked to sulfur metabolism via redox cofactors. This complex could potentially contribute to symbioses, for example with sulfur-oxidizing bacteria such as Thiothrix that is based on cryptic sulfur cycling. One Doudnabacteria genome encodes adjacent sulfur dioxygenase and rhodanese genes that may convert thiosulfate to sulfite. We find similar conserved genomic architecture associated with CPR bacteria from other sulfur-rich subsurface ecosystems. Our combined metagenomic, geochemical, spectromicroscopic and structural bioinformatics analyses of biofilms growing in sulfide-rich springs revealed consortia that contain CPR bacteria and sulfur-oxidizing Proteobacteria, including Thiothrix, and bacteria from a new family within Beggiatoales. We infer roles for CPR bacteria in sulfur and hydrogen cycling.

***N.B**. All main figures and tables for this manuscript can be found below in sections 1.6 and 1.7. All supplementary files (including figures and tables) can be found online with the published manuscript.*

1.1 Introduction

Sulfur is the fifth most abundant element on earth and the sulfur cycle is a key component of Earth's interlinked biogeochemical cycles (Sheik et al. 2015; Anantharaman et al. 2014). In natural ecosystems, sulfur exists in several oxidation states, -2, 0, +2, +4 and +6 being the most common, in the forms of polysulfide (HS_x or S_x^{2-}; -2,0), thiosulfate ($S_2O_3^{2-}$; -1,+5), tetrathionate ($S_4O_6^{2-}$; -2,+6), sulfite (SO_3^{2-} ; +4) and sulfate (SO_4^{2-} ; +6). Microbes play an important role in sulfur cycling in aqueous and soil environments. H_2S is a toxic compound that must be maintained at low levels for the sustained growth of microbial consortia, thus microbial sulfide oxidation is beneficial at the community level.

Sulfide (S^{2-}) is common in natural springs and can serve as a source of energy and reducing power for chemolithoautotrophic microorganisms. Chemolithoautotrophic microbial communities with members that carry out the oxidation, reduction and disproportionation of sulfur compounds are found in environments such as hydrothermal vents (Kalanetra et al. 2004; Takai et al. 2005), water column oxic/anoxic interfaces (Yakushev et al. 2007; Grote et al. 2012; Madrid et al. 2001), terrestrial caves (Macalady et al. 2006; Macalady et al. 2008; Engel 2004), groundwater (Sarbu et al. 1994; Sharrar et al. 2017) and activated sludges (Williams and Unz 1985). *Beggiatoaceae* and *Thiotrichaceae* that have been cultivated have been shown to use hydrogen sulfide either mixotrophically or heterotrophically (Nelson et al. 1986; Nielsen et al. 2000; Hinck et al. 2007; Ehrlich and Newman 2008). *Beggiatoa* spp. are gliding filamentous bacteria that form intracellular spherical S^0 granules that may oxidize to sulfate when H2S supply becomes limited (Sweerts et al. 1990). *Thiotrix spp*. are gliding bacteria that can grow as long multicellular filaments (cells in a microtubular sheath), can form rosettes, and are known to accumulate intracellular spherical S^0 granules when in the presence of reduced sulfur compounds (Williams and Unz 1985; Rossetti et al. 2003) and organics or CO_2 (carbon and energy source) (Nelson et al. 1986). Prior work (Inagaki et al. 2004; Jones et al. 2008; Hamilton et al. 2014; Rossmassler et al. 2016; Meziti et al. 2021) indicates that sulfur-oxidizing bacteria support communities by providing resources such as fixed carbon and nitrogen.

To date, most studies of sulfur-based chemoautotrophic ecosystems have investigated the roles of the relatively most abundant organisms. However, it is well understood that microbial biofilms are structured as networks of interacting organisms, some of which are fundamentally dependent on other community members. Of particular interest are CPR bacteria (also known as Patescibacteria) (Wrighton et al. 2012; Brown et al. 2015; Hug et al. 2016; Parks et al. 2018) that can form symbioses with host organisms (He et al. 2015; Bor et al. 2019; He et al. 2021). Prior surveys have documented CPR bacteria in sulfur-based communities (Wrighton et al. 2012; Vigneron et al. 2021; Hahn et al. 2022), yet the nature of CPR-host relationships and the roles of CPR in sulfur-based communities remain under-explored.

Here, we studied chemoautotrophic microbial communities sustained by sulfur metabolism in two mineral springs MS4 and MS11 at Alum Rock Park, CA, USA, where sulfide-rich groundwater discharges along the Hayward fault. We profiled oxygen isotopes, temperature, water composition and spring discharge rates to constrain the sources of water. We then further combined genome-resolved metagenomics with X-ray spectromicroscopy and scanning electron microscopy to investigate metabolic capacities, interdependencies, and structure of the microbial biofilm community at these two springs. Synchrotron-based spectromicroscopy revealed a close association between ultra-small cells and sulfur-oxidizing bacteria. The possibility that these bacteria are CPR might indicate the existence of a cryptic chemoautotrophic ecosystem. We predict the contributions of the major community members to carbon, nitrogen, sulfur and hydrogen cycling and investigate the potential roles of the abundant and diverse CPR bacteria in these consortia.

1.2 Materials and Methods

Site Description and Microbial biomass collection

The spring system is located along Penitencia Creek in Alum Rock Park, San Jose, CA (37°23'57.7"N, 121°47'48.8"W) (Figure 1.1A**)**. The two sample sites, Mineral Springs 4 and 11 (MS4 and MS11) are located on opposite sides of the creek approximately 250 m from one another (Figure 1.1B-C). Samples for geochemical analyses were taken in May 2005, during the dry season, and were filtered on-site using sterile 0.2 µm filters. Biofilm samples for scanning electron microscopy were collected from both sites using sterile pipettes. Samples were then transported back to the laboratory on ice, and solutions for cation analyses were acidified with 3% nitric acid. Biofilm samples for metagenomic sequencing were collected on , June 13, 2015, July 2, 2019 and July 24, 2020. Planktonic samples were collected on June 10, 2015 and July 24, 2020. Two sets of planktonic samples were taken by sequentially filtering 379 L and 208 L of water, respectively, from the MS4 spring onto 0.65 µm and 0.1 µm large volume filters (Gravertech 5 inch ZTEC-G filter). Filters were frozen on dry ice at the site and stored at -80°C for genome-resolved metagenomic analyses. For synchrotron-based measurements (STXM and X-ray microprobe), thin white streamers were collected on June 13, 2015 with sterile tweezers at both sites and flash-frozen on site. A 2 µL droplet of biofilm sample was deposited using a sterile pipette onto either a 3 mm diameter Si_3N_4 window (SiMPore) or a TEM Cu-grid (300 mesh, lacey carbon coated formvar, Ted Pella Inc.) and were manually blotted with filter paper (Grade 1 filter paper, Whatman®) and immediately plunged into liquid nitrogen using a portable LN_2 plunger, gas ethane (used for flash-freezing) was not available at the time of sampling. Samples were not rinsed or spinned so as to preserve the structural integrity of the filaments and preserve the CPR bacteria-bacteria-filaments spatial relationships. Samples were stored in a LN_2 storage dewar (Taylor-Wharton, 34L) until measurements.

Geochemical Analysis

Water discharge (volume/time) was measured by diverting water into either a bucket or graduated cylinder to measure volume, and time was recorded with a stopwatch. Temperature was measured with a type K thermocouple until February 2008 and thereafter with a thermistor. Accuracy is 0.2 ^{o}C and 0.1 ^{o}C, respectively. Water for O and H isotope measurements was collected in 250 mL Nalgene bottles. Discharge and temperature were not measured if outflow channels from the springs backed up to create pools of water. Cation analysis was performed on a PerkinElmer 5300 DV optical emission ICP with autosampler. Anion analysis was performed on-site using a HACH DR2010 spectrophotometer with protocols provided by the manufacturer. O and H isotopes were measured with a GV IsoPrime gas source mass spectrometer, with analytical precision of approximately 0.1 and 1 per mil, respectively.

Scanning Electron and Confocal Fluorescence Microscopy

Biofilm samples for Scanning electron microscopy were fixed for two hours in a 2% glutaraldehyde solution (in 0.1 M sodium cacodylate buffer) according to a standard protocol, then vacuum aspirated onto 0.22 µm polycarbonate filters (Osmonics, poretics, 47 mm, Catalog number K02CP04700), and rinsed three times in 0.1 M sodium cacodylate buffer. The samples were then dehydrated in successive ethanol baths of increasing concentration and dried using a Tousimis AutoSamdri 815 Critical Point Dryer for approximately one hour. Specimens were mounted on gold stubs and sputter coated with a gold/palladium mix. Imaging was performed on a Hitachi S-5000 scanning electron microscope at 10 keV at UC Berkeley.

Biofilm samples for confocal fluorescence microscopy were prepared with fluorescence dyes. Cell membranes and nucleic acids in the MS4 and MS11 biofilms were stained by adding simultaneously the lipophilic dye F4-64 (20 µg/mL) (Thermo-Fisher, Grand Island, NY, USA) and SYTOX Blue (20 µg/mL) (Thermo-Fisher, Grand Island, NY, USA), respectively. Samples were incubated at room temperature for 20 minutes, and then a 10 µL aliquot of biofilm was deposited onto a glass slide. Samples were imaged using a Leica Stellaris 5 confocal fluorescence microscope. Images were acquired using a 63x oil immersion objective with laser excitation for SYTOX Blue (480 nm) and FM4-64 (520 nm), keeping the same parameters for both biofilms. Unmixed images were combined and false-colored using the Leica Application Suite X (LAS X). All data was collected at the Innovative Genomics Institute, UC Berkeley.

Scanning Transmission X-ray Microscopy (STXM)

STXM and near edge x-ray absorption fine structure (NEXAFS) spectroscopy measurements were performed on the soft X-ray undulator beamline 11.0.2 (Kilcoyne et al. 2003) of the Advanced Light Source (ALS), Berkeley, CA, USA. Data were recorded with the storage ring operating in top-off mode at 500 mA, 1.9 GeV. Frozen samples were thawed right before STXM-NEXAFS measurements at ambient temperature under He at pressure <1 atm. A Fresnel zone plate lens (40 nm outer zones) was used to focus a monochromatic soft X-ray beam onto the sample. The sample was raster-scanned in 2D through the fixed beam and transmitted photons were detected with a phosphor scintillator-photomultiplier assembly; incident photon counts were kept below 10 MHz. The imaging contrast relies on the excitation of core electrons by X-ray absorption (Stöhr 1992; Ade et al. 1992; Kirz et al. 1995). STXM images recorded at energies just below and at the elemental absorption edge (S L_3 and C K) were converted into optical density (OD) images where the OD for a given energy can be expressed from the Beer-Lambert law as OD= $-\ln(I/I_0)= \mu\, \rho\, t$, where I, I_0, µ, ρ and t are the transmitted intensity through the sample, incident intensity, mass absorption coefficient, density and sample thickness, respectively. Protein, carbon and elemental sulfur maps were obtained by taking the difference of OD images at 280 and 288.2 eV, at 280 and 305 eV, and at 162 and 163.9 eV respectively. Image sequences ('stacks') recorded at energies spanning the S $L_{2,3}$-edges (160-180 eV) with steps of 0.3 eV around the L_3-edge, and C K-edge (280-305 eV) with steps of 0.12 eV around the K-edge were used to obtain NEXAFS spectra from specific regions. S $L_{2,3}$ edges NEXAFS spectra are affected by spin-orbit coupling (multiplet structure) and provide information on the oxidation state of sulfur. Beam-induced radiation damage was carefully checked.

Additionally, STXM-NEXAFS measurements at 87 K were performed on frozen-hydrated samples deposited on Si_3N_4 windows so as to preserve sample morphology and chemical integrity (Comolli and Downing 2005) and minimize beam-induced radiation damage[40]. We made sure to analyze exclusively fully frozen-hydrated regions of the samples by first imaging the entire window. The samples were cryo-transferred through a specimen chamber (<100 mTorr) into an LN_2-cooled stage (87 K) inside the STXM operated with a scanning Fresnel zone plate lens (60 nm outer zones), under vacuum (10^{-6} torr). With this setup, the sample is not rastered-scanned through the fixed beam so as to minimize sample vibrations; instead the zone plate is scanned in 2D. Note that sulfur $L_{2,3}$ -edges could not be accessed with this setup due to geometrical constraints.

At least two different sample regions were analyzed at each elemental edge. The theoretical spectral and spatial resolutions during measurements were +/-100 meV ; 40 nm (at room temperature) and 60 nm (at 87 K) respectively. The photon energy was calibrated at the C K-edge using the Rydberg transition of gaseous CO_2 at 292.74 eV (C 1s→ 3s ($\nu = 0$)). Sulfur spectra were calibrated using the S $2p_{3/2}$ edge of elemental sulfur set at 163.9 eV. An elemental sulfur standard spectrum was kindly provided by Geraldine Sarret (Sarret et al. 1999) (University Grenoble Alpes). A lipid standard compound (PE) was kindly provided by Susan Glasaeur (University of Guelph). Carbon spectra were pre-edge background subtracted using a linear function and normalized at 300 eV. All data was processed with the aXis2000 software version 06 Jul 2021 (http://unicorn.mcmaster.ca/aXis2000.html). Sulfur spectra were further pre-edge background subtracted and post-edge normalized using Athena version 0.9.26 (Demeter package (Ravel and Newville 2005)).

X-ray fluorescence microprobe (XFM)

Synchrotron XFM measurements were performed in cryogenic conditions (95 K) at ALS XFM beamline 10.3.2 (Marcus et al. 2004), with the storage ring operating in top-off mode at 500 mA, 1.9 GeV. Micro-focused X-ray fluorescence (µXRF) elemental mapping was performed on LN_2-frozen hydrated samples oriented at 45° to the incident X-ray beam, frozen biofilm samples were mounted onto a TEM cartridge and cryo-transferred into a LN_2-cooled apparatus following methods described elsewhere (Fakra et al. 2018). All data were recorded using a single-element XR-100 silicon drift detector (Amptek, Be window).

XRF maps were recorded at 4138 eV (100 eV above the Ca K-edge) using a beam spot size of 3 µm x 4 µm, 2 x 2 µm pixel size and 70 ms dwell time/pixel. Micro-XRF spectra were recorded simultaneously on each pixel of the maps. All maps were then deadtime-corrected and decontaminated using custom LabVIEW 2018 (National Instruments, Austin, TX, USA) software available at the beamline. Maps were then processed using a custom Matlab R2020b program (MathWorks, Natick, MA, USA) available at the beamline.

DNA extraction and metagenomic sequencing

Approximately 200 µl of biofilm was extracted using MoBio Powersoil DNA extraction kit (MoBio Laboratories, Inc., CA, USA) according to the manufacturer's protocol, with the bead-beating time reduced to less than one minute. This DNA extract was then gel purified and quantified using a low-mass ladder (Promega).

Total genomic DNA for metagenomic sequencing (150 bp or 250 bp reads) for both biofilm and planktonic samples (20% of each filter) was extracted using MoBio PowerMax Soil DNA extraction kit. Cells were extracted from 20% of each filter by adding 15 ml of lysis buffer and vortexing for 10 minutes. Lysis of cells was modified by heating to 65°C for 30 minutes and 1 min of bead beating. DNA was eluted in milliQ water and ethanol precipitation was performed (70% EtOH, 3 M sodium acetate, incubation for 24 hours at 4 ˚C).

Illumina sequencing, assembly, binning and sequence curation

Shotgun genomic reads were assembled using IDBA-UD (Peng et al. 2012). Contigs larger than 2.5 kb were retained, and sequencing reads from all samples were mapped against each resulting assembly utilizing Bowtie2 (Langmead and Salzberg 2012). Differential coverage profiles, filtered with a 95% read identity threshold, were then used for genome binning using a suite of binning tools (MetaBAT2 (Kang et al. 2019), VAMB (Nissen et al. 2021), MaxBin2 (Wu et al. 2016), Abawaca (https://github.com/CK7/abawaca), with the final bin choice determined by DAS Tool (Sieber et al. 2018). Draft genomes consisting of scaffolds ≥ 1 kbp in length were binned using ggKbase manual binning tools based on a combination of GC content, coverage, single copy gene content, phylogenetic profile and patterns of organism abundance over samples. The phylogenetic profile was established using representative genomes from the GTDB database. In some cases, scaffold sequences from groups of bins were used to construct emergent self-organizing maps in which the structure was established using tetranucleotide composition (tetra-ESOMs). For scaffolds > 6 kb, scaffolds were subdivided into 3 kb segments and treated separately in the ESOM analysis. In cases where the majority of segments from the same scaffold did not group together in the ESOM, the scaffolds were evaluated manually (based on gene content and other information) to resolve their placement or assign them to unbinned. The scaffold set defined based on ESOM analysis was then used to generate a draft genome bin that was again checked for consistent binning signals (as above). As ESOMs only used scaffolds >3 kb in length, scaffolds from the original bins were added if they had a tightly defined GC, coverage and the expected phylogenetic profile. CheckM2 (Chklovski et al. 2023) was used for estimation of genome completeness, strain heterogeneity and contamination. Curated genomes with less than 5 duplicated single-copy genes (some of which occur because genes are split at scaffold ends) and with ≥ 95% of the expected single copy marker gene were used for completeness estimation: at least 42 of 51 single copy genes used for preliminary bin evaluation in ggKbase were classified as near-complete. Genomes with >5 duplicated single-copy genes were classified as partial, regardless of other indicators of bin completeness.

Phylogenetic analyses

The concatenated ribosomal protein tree was generated using 16 syntenic genes that have been shown to undergo limited lateral gene transfer (rpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, 24 and rpS3, 8, 10, 17, 19) (Sorek et al. 2007). We obtained branch support with the ultrafast bootstrap (Minh et al. 2013) implemented in iQ-TREE v1.6.12 (Nguyen et al. 2015) with the following parameters: -bb 1000 -m LG+F+G4. Trees were visualized using iTOL v6.3.2 (Letunic and Bork 2021). Amino acid alignments of the individual ribosomal proteins were generated using MAFFT v7.304 (Katoh and Standley 2013) and trimmed using trimAL (Capella-Gutiérrez et al. 2009) with the following setting: -gt 0.1.

To verify the presence of biogeochemically-relevant genes, phylogenetic trees were constructed. We used gene markers for sulfur (DsrAB, Pdo), carbon metabolism (RuBisCO) and energy conservation ([NiFe]-hydrogenases). Sequences were obtained using GOOSOS and aligned using MAFFT v7.304. All other phylogenies were generated using iQ-TREE v.1.6.12 using the ultrafast bootstrap and parameters specified previously.

Hydrogenase sequences from Alum Rock genomes were obtained using HMMs (Matheus Carnevali et al. 2019). The phylogenetic classification was performed using reference sequences obtained from (Matheus Carnevali et al. 2019) and HydDB (Søndergaard et al. 2016). Verification of hydrogenase loci was performed via inspection of nearby genes and the presence of required hydrogenase accessory genes. Genome context diagrams were generated using Clinker (Gilchrist and Chooi 2021).

Metagenomics metabolic pathways analysis

Preliminary functional annotations were established using METABOLIC (Zhou et al. 2020) and collections of metabolic capacities in genome bins were overviewed using ggKbase tools (Raveh-Sadka et al. 2015). In addition, metabolic profiling was done by mapping ORFs to KEGG ortholog groups (KOs) using an HMM database that was compiled as previously described (Castelle et al. 2018). This HMM database was used to scan the metagenomic bins, and ORFs were assigned the KO of the best-scoring HMM, providing it was above the noise threshold. In addition, we profiled metabolic capacities with KEGG functional annotation using METABOLIC (Zhou et al. 2020).

Protein structure prediction

Protein structures were predicted for the putative complexes of the nitrate reductase (Nrx), dioxygenase/rhodonase, and group 3b [NiFe]-hydrogenase using AlphaFold2 in multimer mode. Specifically, for the 3b [NiFe]-hydrogenase complexes, AlphaFold2 was used in multimer mode for the HyhL (hydrogenase large subunit), HyhS (hydrogenase small subunit), HyhG (diaphorase catalytic subunit) and HyhB (diaphorase electron transfer subunit) (Jumper et al. 2021; Evans et al. 2022). In all cases, the average per residue confidence scores (pLDDT) exceeded 90, a level that is empirically shown to produce highly accurate local structural models. The best-scoring models were aligned to related protein complexes in PyMol.

1.3 Results

Groundwater of mixed origin hosts biofilms dominated by filamentous bacteria

We measured the flow rate, pH, and concentrations of ionic species (Supplementary Table S1) in the MS4 and MS11 groundwater. The MS11 spring has a higher flow rate, ionic strength, alkalinity, and sulfide levels than the MS4 spring. H and O stable isotope compositions of the waters, combined with salinity measurements, indicate that spring waters are mixtures of meteoric input and pore waters from the host Miocene Monterey Group shales and cherts, and possibly deeper Cretaceous sediments of the Great Valley Group. MS4 water is more diluted by meteoric input than MS11. Long-term monitoring of these two springs shows they experience small seasonal fluctuations in temperature and that they are generally hydrologically and geochemically stable (Figure 1.1D-F). Water temperatures of 27-29 ^{o}C are well above the mean annual surface temperature of 15.1 ^{o}C. The salinity of the springs is 1.8 and 2.3% for MS4 and MS11, respectively. The sulfide levels (within the zone of oxygenation) range up to ~9 and 69 µmol/L at MS4 and MS11, respectively.

The biofilms at both MS4 and MS11 sites (Figure 1.1A) are mainly composed of thin white streamers (~ 5-10 cm long) that are primarily attached to rocks. Scanning electron microscopy (SEM), confocal fluorescence microscopy (CFM) and scanning transmission X-ray microscopy (STXM) revealed that MS4 biofilms consist of filaments and cells distributed amongst the filaments (Figure 1.2; Figure 1.3) whereas the MS11 biofilm consists mainly of filamentous bacteria.

Filamentous bacteria have encapsulated elemental sulfur granules and episymbionts

Sulfur µXRF distribution maps at 95 K evidenced the presence of sulfur across MS4 filamentous bacteria. STXM sulfur maps (Figure 1.2) and S $L_{2,3}$ NEXAFS spectra showed that the filaments contain spherical S^0 granules encapsulated in protein-rich compartments (Figure 1.3A-B). The spherical granules in the middle filament (Figure 1.2E) are roughly 380 nm average diameter, as estimated from 76 granules. The width of these MS4 filaments is <1.6 µm. Filaments exhibit septa and longitudinal cell envelopes as observed by STXM (Figure S4) and CFM (Figure 1.2C). Rod-shaped, curved-shaped, and few coccoid-shaped cells were found near the filaments in MS4 biofilms, as well as ultra-small cells (Figure 1.2), often found within extracellular polymeric substances (EPS). Higher-resolution protein maps of MS4 and MS11 filaments (Figure 1.3) suggest that sulfur granules are surrounded by proteins, as further confirmed on maps recorded at 87 K (Figure 1.4).

Carbon K-edge NEXAFS spectra at 87 Kelvin of filamentous bacteria in MS4 and MS11 biofilms (Figure 1.4) exhibit similar spectra, with a major peak at 288.2 eV (amide carbonyl groups in proteins, (Stewart-Ornstein et al. 2007)), a peak at 285.2 eV attributed mainly to aromatic groups in proteins and peaks at 286, 286.6, 287.4 and 289.4 eV that can be attributed to nucleic acids (Boese et al. 1997; Samuel et al. 2006; Zubavichus et al. 2008) (Figure S5). One prior report (Zubavichus et al. 2008) suggests that spectra of nucleotide bases with peaks occurring within 284.7−286.9 eV are likely associated with $\pi^*(C=C)$, whereas peaks in the 287.3−287.4 eV range are likely associated with $\pi^*(C=N)$. Resonances are more defined than in prior studies at room temperature (Ade et al. 1992; Benzerara et al. 2004; Toner et al. 2009; Fakra et al. 2018), due to reduced Debye-Waller thermal disorder at low temperatures. This in turn allows us to unequivocally detect the presence of nucleic acids in the filaments, in addition to proteins. Copious extracellular polymeric substances surrounding MS4 and MS11 filaments exhibit a main peak at 288.7 eV, which can be attributed to carboxyl groups in acidic polysaccharides (Supplementary Table S2). The spectra of MS11 cells are also similar to those of filamentous bacteria, with a major peak at 288.2 eV (amide bonds) and other peaks associated with nucleic acids. The spectrum of a surface-attached cell on an MS4 filament exhibits a major peak at 288.2 eV (amide bonds) and nucleic acid-associated peaks, the 288.2 and 285.2 eV peaks are broader likely due to the presence of EPS surrounding this cell. Cells, filaments and EPS all exhibited a shifted carbonate peak at 290.7 eV that corresponds to either organic carbonates or carbonate minerals (Brandes et al. 2010), and originates mainly from dissolved carbonates and carbonate precipitates present in the groundwater at circumneutral pH (Supplementary Tables S1 and S2).

Strikingly, small cells were found along the surfaces of the filaments in MS4 (Figure 1.2F-G) and MS11 biofilms, these cells are typically about 480 nm long, 250 nm wide, as estimated from STXM images. Ultra-small cells (290 ±20 nm long, 120 ±15 nm wide) were also found in close proximity to the filaments in MS4.

Biofilms contain diverse bacteria including CPR bacteria

To determine the bacterial community composition across biofilm samples collected at MS4 and MS11 sites over the years, we analyzed the relative abundance of bacterial taxa at the Phylum, Class, and Order levels. We observed a diverse range of bacterial phyla across the samples (Figure 1.1G-H, see also supplementary material for details). The most abundant phylum across the samples was Campylobacterota, representing >50% of the total bacteria, followed by Proteobacteria and Desulfobacterota, respectively. Data on the MS11 biofilms in the year 2020, indicate a possible shift in the bacterial community structure during this period. A closer look at the class level revealed that Gammaproteobacteria are predominant in most samples from the MS4 site, representing a significant portion of the bacterial community. However, distinct patterns of class-level diversity were evident in different samples, showcasing the dynamic nature of biofilm communities over the years at both sites.

We used genome-resolved metagenomics to investigate microbial consortia, metabolisms and microbial interactions that underpin the Alum Rock communities. In total, we recovered 212 non-redundant genomic bins from the MS4 and MS11 samples (57 from MS11 and 155 from the biofilm + planktonic samples from MS4). Of these, 38 were classified as near-complete (>95%). Taxonomic affiliations of all of the bacterial genomes were established based on concatenated ribosomal protein trees (Figure 1.5A). Genomically represented groups in the biofilms and planktonic fractions from both sites include Gammaproteobacteria (Thiotrichales, Chromatiales, Beggiotales), Campylobacterota (Campylobacterales), Betaproteobacteria (including *Thiomonas*), Deltaproteobacteria (specifically Desulfobacterales), Bacteroidota, Chloroflexota, Ignavibacteria, Spirochaetes, Lentisphaerae, Riflebacteria, Verucomicrobia, Acidobacteria, KSB1, Caldisericota, Planctomycetota, Edwardsbacteria, Dependentiae (TM6), and Margulisbacteria. Diverse groups of CPR are present, including Uhrbacteria (OP11), Gracilibacteria (BD1-5), Peregrinibacteria (PER), Moranbacteria (OD1), Woesebacteria (OP11), Roizmanbacteria and Gottesmanbacteria (OP11), Saccharibacteria (TM7), Falkowbacteria (OD1), Absconditabacteria (SR1), Berkelbacteria, Doudnabacteria and Dojkabacteria (WS6).

To estimate the relative abundance of organisms in the two springs (independent of binning) we calculated the DNA read coverage of all of the genomic bins for each spring (Figure S6). The MS4 spring is dominated by Halothiobacillales, Beggiatoales, Thiotrichales and Campylobacterales based on relative abundance among genomes. The most abundant species in MS4 shares genome-wide average 51% amino acid similarity with the sulfur oxidizer *Thiothrix nivea* (Lapidus et al. 2011). The MS11 spring is dominated by a single *Beggiatoa* sp. (Beggiatoa-related_37_1401).

Diverse bacteria are implicated in sulfur cycling

We used a list of high-confidence functional gene assignments for all draft or better-quality genomes to resolve the capabilities central to the metabolisms of the dominant bacteria in the springs. Not surprisingly, genes encoding sulfur cycling are common in the most abundant organisms at both sites.

Based on the community composition of MS4 and MS11 biofilms, we further focused our analysis on the metabolic pathways of MS4 bacteria where we detected ultrasmall and surface-attached cells on filamentous bacteria implicated in sulfur oxidation. The most abundant organism in MS4, which is closely related to the filamentous bacterium *Thiothrix nivea*, encodes genes (*soxABC*, periplasmic thiosulfate-oxidizing; *aprAB*, adenylylsulfate reductase; *dsrAB*, reverse dissimilatory sulfite reductase) to convert sulfide to thiosulfate, elemental sulfur and sulfate (Figure 1.5B). The absence of *dsrD* genes indicates that the Dsr complex operates in the sulfide oxidation direction (i.e. rDsr pathway). This *Thiothrix* bacterium also lacks any *soxC* genes, which in bacterial genomes has been associated with the accumulation of sulfur granules or polysulfide (Friedrich et al. 2005; Frigaard and Dahl 2009).

MS4 contains various other bacteria capable of oxidizing reduced sulfur compounds. A subdominant population of *Sulfurovum* bacteria encode *sqr* genes and thus likely oxidize sulfide to S^0. Some *Sulfurovum* bacteria in both communities have genomes that also encode *soxCDYZ* complexes, suggesting they mediate thiosulfate oxidation (potentially coupled to nitrate reduction, e.g., via *narG* and *napA*. *Sulfuricurvum* species are also relatively abundant in MS4 and encode genes for sulfur and thiosulfate oxidation, in line with culture-based studies (Handley et al. 2014). The genomes of *Chloroflexota* encode the capacity for thiosulfate disproportionation via thiosulfate reductase / polysulfide reductase (*phsA*) and sulfide oxidation via flavocytochrome *c* sulfide dehydrogenase. Two low abundance Gammaproteobacteria species related to *Acidthiobacillus* have the capacity for thiosulfate oxidation. Several genomes from moderately abundant *Halothiobacillales* have the metabolic capacity for sulfide and thiosulfate oxidation via *fccB*, *dsrAB* and *soxBCY* respectively.

Some bacteria from MS4 spring also potentially mediate dissimilatory sulfate reduction. Specifically, the genomes of some *Desulfobacteriales* belonging to the families of *Desulfatiglandaceae*, *Syntrophobacterales, Desulfurivibrionaceae* and *Desulfarculales* encode the capacity to reduce sulfate back to sulfide via Dsr genes, likely coupled to oxidation of organic carbon or H_2. Some rare Desulfocapsaceae from MS4 that are related to bacteria of the genus *Desulfocapsa* have thiosulfate reductase, group 3b [NiFe] (Hyd; possibly sulfhydrogenase), as well as SAT (Sulfate adenylyltransferase) and APR (Adenylylsulfate reductase) for the oxidation of sulfite to sulfate. Thus, it appears these bacteria are involved in sulfur disproportionation whereby S^0, thiosulfate, and sulfite are converted to H_2S and sulfate, as has been demonstrated in cultures of bacteria from this genus (Finster et al. 1998). Other *Desulfocapsa* spp. have tetrathionate reductase genes, suggesting they are capable of converting tetrathionate to thiosulfate. The *Desulfocapsa*-related bacteria also contain *dsrABD* genes, which fall within the reductive cluster closely related to those from *Desulfocapsa sulfexigens*. We infer that the *Desulfocapsa*-related bacteria are capable of S disproportionation, as previously reported (Finster et al. 2013) and the presence of the *dsrD* functional marker protein suggests that these species in the springs are capable of both S disproportionation and sulfite reduction. Only members of the candidate phylum Riflebacteria, family Ozemobacteraceae, have the capacity of anaerobic sulfite reduction via anaerobic sulfite reductase system (*asrABC*). A bacterium from a new class of *Caldithrix* from the MS4 spring is predicted to perform sulfur oxidation via dissimilatory sulfite reductase, sulfite oxidation, sulfate reduction and thiosulfate disproportionation (Supplementary Tables S6A). We also identified abundant bacteria from novel families of Bacteroidetes, which generally encode thiosulfate reductase genes (*phS*) and adenylylsulfate reductase (*aprA*) involved in thiosulfate disproportionation and sulfate reduction.

Surprisingly, we identified persulfide dioxygenase (sdo) and rhodonase (thiosulfate sulfurtransferase) in the genomes of *Elusimicrobia*, *Riflebacteria*, *Oscillatoriophycidae* and in a novel family of *Syntrophales* (Figure 1.6A). These enzymes are also present in some heterotrophic bacteria, where they play important roles in the detoxification of intracellular sulfide and sulfur assimilation respectively (Motl et al. 2017; Zhang et al. 2020). We also found a putative sulfur dioxygenase encoded in a Doudnabacteria genome that clusters with protein sequences of other CPR bacteria from public data. In the operon, a sulfur transferase is adjacent, suggesting its potential function in thiosulfate oxidation (Figure 1.6B). This is interesting because persulfide dioxygenase has not been previously linked to CPR bacteria. Modeling of the persulfide dioxygenase from Doudnabacterium using AlphaFold2 indicates that it has structural homology with the biochemically characterized persulfide dioxygenase (Figure 1.6B-D). Furthermore, we identified these two adjacent genes in the genomes of several other CPR from high sulfide environments, including Kaiserbacteria (groundwater from California), Pacebacteria (wastewater), Moranbacteria, and Gracilibacteria (Crystal Geyser aquifer). Thus, we suggest that these genes may enable a variety of CPR bacteria to grow and generate energy from sulfur oxidation.

In contrast, the most highly sampled genomes in MS11 spring are from a *Beggiatoa* species including a novel family within this group (Supplementary Tables S6A). As expected, these *Beggiatoa* genome encodes the Dsr genes (*dsrABCHJKMOPR)*; *dsrD* genes were not identified, we conclude that the Dsr genes are operational in a reverse Dsr pathway (rDsrABs) (Anantharaman et al. 2018). The genome also encodes AprAB (adenylylsulfate reductases), and Sat (sulfate adenylyltransferase) for the oxidation of sulfide to sulfate, sulfide-quinone oxidoreductase (Sqr) as well as sulfide dehydrogenase (*fccB*) genes for the oxidation of hydrogen sulfide to S^0. The genomes contain a partial set of sulfur-oxidizing sox pathway genes, but *soxDXYZ* were identified. Given the lack of *soxC*, we conclude that like *Thiothrix* the primary role of *Beggiatoa* in the community is the conversion of sulfide to thiosulfate, elemental sulfur and sulfate. The absence of *soxCD* in bacterial genomes has been previously associated with the accumulation of sulfur granules or polysulfide (Friedrich et al. 2005; Frigaard and Dahl 2009).

Sulfur-oxidizing bacteria also contribute to nitrogen cycling

The dominant bacteria in MS4 and MS11 springs are predicted to mediate nitrogen fixation and denitrification processes. In both MS4 and MS11, genes encoding nitrogenase implicated in N_2 fixation are widespread in Proteobacteria, including in the dominant *Thiothrix*, *Beggiatoa* and *Sulfurovum* and Verrucomicrobia. Other organisms with this capacity include other *Gammaproteobacteria*, *Chromatiales*, *Campylobacteriales*, *Sulfuricurvum*, *Ignavibacteria*, *Sulfosprillum*, *Spirochaetes*, *Desulfocapsa*, and potentially *Lentisphaerea*.

The *Thiotrichales* genomes encode numerous genes for the reduction of nitrate and nitrite, however only the dominant *Thiothrix* species have the capability to reduce nitrite to nitrous oxide via *nirS* and *norBC* genes. Some *Chromatiales* bacteria at both springs also appear to be capable of dissimilatory nitrite oxidation to ammonia. The sulfur-oxidizing Campylobacterales that we identified in both MS4 and MS11 springs have numerous genes implicated in the reduction of nitrate (*napAB*) and nitric-oxide (*norBC*). Two low abundance *Acidithiobacillales* in MS4 that are predicted to perform thiosulfate oxidation have ammonia monooxygenase (*amoA*) genes, suggesting they may be involved in ammonia oxidation and nitrite ammonification. *Chloroflexota* populations at both springs have the capacity for nitrite reduction via nitrite reductase (*nirK*), nitric oxide reduction (*norBC*) and nitrite ammonification. A novel *Caldithrix* species from MS4 has the potential of nitric oxide reduction via the nitric oxide reductase *norBC* and nitrite reduction via the periplasmic nitrate reductase NapA (Figure 1.5B).

In addition to being the most abundant sulfur oxidizers in the MS11 spring, *Beggiatoa* are metabolically versatile with regards to nitrogen cycling. Their genomes encode genes with similarity to nitrate reductase (*narABG*), nitrite reductase (*nirS*), nitric oxide reductase (*norBC*), and nitrous-oxide reductase (*nosZ*) for the complete reduction of nitrate to N_2. They also contain *nrfA* potentially for dissimilatory nitrite reduction to ammonia (DNRA) or nitrite ammonification. Thus these bacteria can likely couple sulfur oxidation to nitrate reduction, in line with prior studies (Kalanetra et al. 2004; Sharrar et al. 2017).

Extensive links between hydrogen and sulfur metabolism

To gain insight into the role of hydrogen metabolism in the Alum Rock springs, we analyzed the distribution of hydrogenases and associated enzymes in the genomes. There was considerable capacity for fermentative H_2 production using nicotinamides (via group 3b and 3d [NiFe]-hydrogenases), ferredoxin (via group A [FeFe]-hydrogenases and group 4 [NiFe]-hydrogenases), and formate (via formate hydrogenases) as electron donors (Figure 1.7A). Some putative H_2

producers are likely to be metabolically flexible bacteria such as *Sulfurospirillum* and Flavobacteriales, which can switch to fermentation when limited for respiratory electron acceptors based on previous reports (Berney et al. 2014; Søndergaard et al. 2016). CPR bacteria, TA06, and Spirochaetes with group 3b and 3d [NiFe]-hydrogenases are likely to be obligate fermenters given they apparently lack terminal reductases (Supplementary Table S7). The gene arrangements of the group 3b [NiFe]-hydrogenases in the genomes of the CPR bacteria *Berkelbacteria* and *Moranbacteria* (Figure 1.7B) are similar to the biochemically characterized hydrogenase and sulfhydrogenase of *Pyrococcus furiosus* (Pedroni et al. 1995) and those previously reported in other CPR bacteria (Wrighton et al. 2012; Jaffe et al. 2020), suggesting that these hydrogenases may be capable of reversible oxidation of hydrogen or capable of reducing sulfur compounds like polysulfide. We modeled the complex from *Berkelbacteria* genome using AlphaFold, this model suggests a hydrogenase module (α and γ subunits) with an electron wire of FeS clusters connecting to a nucleotide reducing module (β subunit) (Figure 1.7C). The δ subunit has no close structural analogues but contains an additional FeS cluster and may accommodate an additional electron-accepting partner (Figure 1.7D). Based on this structural analysis there are two separate paths for the electrons suggesting this 3b [NiFe]-hydrogenase complex is potentially an electron-bifurcating hydrogenase.

Numerous bacteria in the Alum Rock springs are predicted to consume H_2 for energy generation. Most of these hydrogenotrophs are predicted to use H_2 to reduce sulfate (via group 1b and 1c [NiFe]-hydrogenases; primarily Deltaproteobacteria), elemental sulfur (via group 1e [NiFe]-hydrogenases; primarily Gammaproteobacteria), or heterodisulfides (via group 3c [NiFe]-hydrogenases; various lineages including Acidobacteria). The most abundant Gammaproteobacteria and Campylobacteria likely oxidize both H_2 and sulfur compounds either mixotrophically or alternatively autotrophically. The hydrogenase repertoire of these organisms includes the oxygen-tolerant group 1b and 1d [NiFe]-hydrogenases (Olson and Maier 2002; Fritsch et al. 2011).

Organic carbon cycling and fermentation

The ability to fix inorganic carbon (CO_2) is a common predicted capacity for bacteria from both sites (Supplementary Table S6A and S6B). The dominant *Thiothrix*, *Beggiatoa*, and *Chromatiales*-related bacteria have type II RuBisCO genes that function in the Calvin-Benson-Bassham (CBB) cycle. One Absconditabacteria genome has a RuBisCO that phylogenetic analysis places within the form II/III CPR clade, as reported previously (Wrighton et al. 2012; Wrighton et al. 2016); these enzymes are inferred to function in a nucleoside salvage pathway in which CO_2 is added to ribulose-1,5-bisphosphate to form 3-phosphoglycerate (Sato et al. 2007). *Elusimicrobia* and *Campylobacterota*, including species related to Sulfurimonadaceae, have ATP citrate lyase genes that encode the critical enzyme for CO_2 fixation via the reverse TCA (rTCA) cycle. We also identified rTCA genes in a novel *Bacteroidetes* organism (Supplementary Table S6A and S6B). Genes of the Wood Ljungdahl carbon fixation pathway (*cooS/acsA, acsB and acsE*) were widespread in both springs, including in members of the Bacteroidetes, Desulfocapsa, Lentisphaerae, Chloroflexi, and Aminicenantia with the potential of oxidation of small organic compounds.

We used marker genes involved in carbohydrate metabolism to infer polymer biomass degradation capacity of the biofilm organisms. Many bacteria in both springs have the capacity to hydrolyze complex organic molecules to produce a variety of electron donors such as acetate, hydrogen, and lactate (Figure 1.9). Of the organisms in the community, *Bacteroidetes* and Ignavibacteria contain the most glycosyl-hydrolase genes and thus they likely play important roles in polysaccharide degradation. Notably, one *Bacteroidetes* from MS11 has 66 glycoside hydrolase genes. This organism is the only bacterium that appears to be capable of degrading cellulose, hemicellulose, polysaccharides, and monosaccharides. *Gammaproteobacteria*, *Spirochaetes*, *Bacilli*, and *Lentisphaerae* also contain genes for the degradation of a variety of complex carbohydrates, but these genes are at relatively low abundance in the sulfur-oxidizing *Proteobacteria*.

Similarly, many bacteria other than the sulfur-oxidizing *Proteobacteria* (and CPR) have indications of the capacity for beta-oxidation pathway of saturated fatty acids to acetyl-CoA. Many CPR bacteria have a few glycosyl hydrolase genes, which is significant given the scarce indications of other metabolic capacities in these microorganisms. Methane oxidation is predicted to be a capacity of members of Verrucomicrobia, specifically members of the *Methylacidiphilales*. This reaction involves particulate methane monooxygenase (pMMO-ABC), the genes for which were identified and classified phylogenetically.

One of the more interesting organisms present in the MS4 spring is a Gracilibacteria, which is predicted to have minimal metabolic capacities beyond glycolysis, production of peptidoglycan, and generation of formate, which may be available for use by other community members. Other capacities predicted for this bacterium are the production of riboflavin, amino-sugars, RNA degradation, 1C by folate, interconversion of purines and pyrimidines, and biosynthesis of a few amino acids (Figure 1.8).

1.4 Discussion

Some springs are hotspots where resources associated with deeply sourced water can sustain chemoautotrophic ecosystems independent of sunlight. We studied two closely spaced but different sites that discharge a mixture of deeply sourced and shallow groundwater, providing microorganisms with reduced compounds and oxygen. Our research integrated geochemical, synchrotron-based spectromicroscopy, metatranscriptomics and genome-resolved metagenomic data to resolve the network of microorganisms that define the ecosystems. This approach provided insights into organism metabolic capacities, their associations, including those that involve CPR bacteria, and the that contribute to biogeochemical processes that sustain autotrophic ecosystems in the context of their spring-based hydrological setting.

Analysis of the metabolisms of the dominant bacteria in the springs revealed that genes implicated in sulfur cycling are common at both sites (Figure 1.9). As expected, the primary energy source is reduced sulfur in the form of sulfide. Overall, the most common sulfur metabolisms are sulfide oxidation, thiosulfate disproportionation, sulfur oxidation, and less commonly, sulfite oxidation and sulfate reduction. Sulfide can be oxidized aerobically and, in some cases, anaerobically, coupled with nitrate reduction. Our metagenomic analyses suggest that intermediate sulfur compounds and sulfate and sulfide are actively cycled by Campylobacterota (*Sulfurovum, Thiovulum*), Gammaproteobacteria (*Thiotrichales and Beggiotales*) in the spring communities, probably coupled to nitrogen compound reduction in some microhabitats. Elemental sulfur serves as an energy source stored as sulfur granules as observed by STXM, in *Beggiatoa* (Schwedt et al. 2011). Interestingly, elemental sulfur-bearing granules may serve as an energy source for the growth of *Beggiatoa* and/or *Thiotrix*. The sulfur oxidizers are the primary source of fixed carbon and nitrogen in the ecosystem.

A higher flow rate and a higher concentration of sulfate was observed at MS11 compared to MS4, and the communities have distinct microbial community composition. The MS4 ecosystem is highly diverse and dominated by abundant sulfide-oxidizing Gammaproteobacteria (*Thiothrix, Sulfurovum*) and sulfate-reducing Desulfobacteriales. The MS11 spring has relatively low diversity and is highly dominated by Campylobacterota (*Sulfurovum, Thiovulum*) and Gammaproteobacteria (*Thiotrichales and Beggiotales*). Our findings are consistent with predictions from studies that indicate that filamentous Campylobacterota dominate biofilms with high sulfide/oxygen (>150) ratios, whereas Gammaproteobacteria (*Beggiatoa*-like) prefer lower (<75) ratios (Macalady et al. 2008).

We further investigated the metabolic capacities of several CPR bacteria within these communities, as their roles in sulfur-based chemoautotrophic ecosystems remain poorly known. CPR bacteria are often characterized by small genomes and minimal anaerobic fermentative metabolism (Luef et al. 2015), however recent studies have shown auxiliary metabolisms such as the presence of hydrogenases (Wrighton et al. 2012; Jaffe et al. 2020), rhodopsin[91], nitrite reductases (Danczak et al. 2017) and F-type ATPase (Chaudhari et al. 2021), that may contribute to alternative energy conservation and adaptations to different environments and host associations. Notably, we identified genes potentially involved in elemental sulfur reduction (Sulfyhydrogenase) and thiosulfate oxidation (persulfide dioxygenase and rhodonase) in the genomes of several CPR bacteria, suggesting a potential new energy generation mechanism for these bacteria. We found that other CPR bacteria from high sulfur environments have the same predicted potential for thiosulfate oxidation, suggesting an important general adaptation of CPR bacteria in sulfur-rich environments.

The most interesting aspect of the current study regards interactions involving CPR bacteria and their host microorganisms. CPR-host associations have rarely been documented, with the exception of oral microbiome-associated Saccharibacteria (TM7) (He et al. 2015; Cross et al. 2019) and Actinobacteria, further laboratory studies (Tian et al. 2022) have validated genomic predictions of metabolic interdependency (Jaffe et al. 2020). One study suggested the presence of CPR cells on the surfaces of their Actinobacteria hosts via SEM and showed them to be rod-shaped and < 0.2 µm in diameter and ~0.5 µm in length (Bor et al. 2020). Another study linked *Vampirococcus* with anoxygenic photosynthetic Gammaproteobacteria (Moreira et al. 2021). Two studies suggest links between Parcubacteria and archaea, in one case *Methanosaeta[98]* and *Methanothrix[98]*. In the case of the CPR Nealsonbacteria associated with *Methanosaeta,* cryo-TEM imaging indicateds that *Methanosaeta*-attached cells are ~0.5 µm in diameter. Other cultivation-independent studies have verified that CPR cells are ultra-small and can be better analyzed via filtration through a a 0.2 µm pre-filtering (Luef et al. 2015). Cryo-TEM imaginges and tomographic analyses have documented ultra-small cells directly associated with CPR cells and host bacteria (Luef et al. 2015; He et al. 2021). Generally, these data indicate that CPR cells are a fraction of a micron in length and diameter, consistent with the size for filament-associated ultra-small cells reported here (<650 nm long, ~250 nm wide, the smallest being 290 ± 20 nm long, 120 ± 15 nm wide). In the MS4 biofilms, ultra-small cells were found associated with the surfaces of long filamentous bacteria containing relatively large S^0 granules as evidenced by STXM. Given that the only abundant filamentous bacteria in these samples are sulfide-oxidizing *Thiothrix,* we predict that some of these tiny episymbiontic cells are CPR bacteria. CPR identifications include Gracilibacteria, Berkelbacteria, Moranbacteria or Doudnabacteria, based on microbial community abundance information. Unfortunately, our attempts to perform fluorescent in-situ hybridization to determine CPR bacterial identity were unsuccessful due to low amount of material, so this inference remains tentative. Future laboratory co-cultivation of *Thiothrix* and their episymbionts may be required to identify the CPR types, so as to better understand the nature of their association (*e.g.*, mutualistic, parasitic). If confirmed, and given the prediction that some CPR bacteria have putative sulfhydrogenases that may produce H_2S (Ma et al. 1993; Pedroni et al. 1995), these episymbionts may be involved in cryptic sulfur cycling that also involves sulfur-oxidizing bacteria.

Hydrogen is an important resource in many environments (Chapelle et al. 2002), yet little is known about the distribution and importance of hydrogenases in sustaining groundwater microbiomes. The most common chemolithoautotrophs in the Alum Rock spring biofilms are H_2-oxidizing bacteria, which use H_2 as an energy source via the enzyme hydrogenase. Specifically, group 3b [NiFe]-hydrogenases are widely distributed in the genomes of many of the microbial community members. These complexes may mediate hydrogen metabolism or the direct hydrogenation of elemental sulfur to hydrogen sulfide (Ma et al. 1993). Other hydrogenases of the microbial community members are implicated in hydrogen production and oxidation. Together, these findings suggest that most bacteria in Alum Rock springs cycle hydrogen gas and sulfur compounds, reactions that underpin the biology and geochemistry of this ecosystem.

1.5 Figures

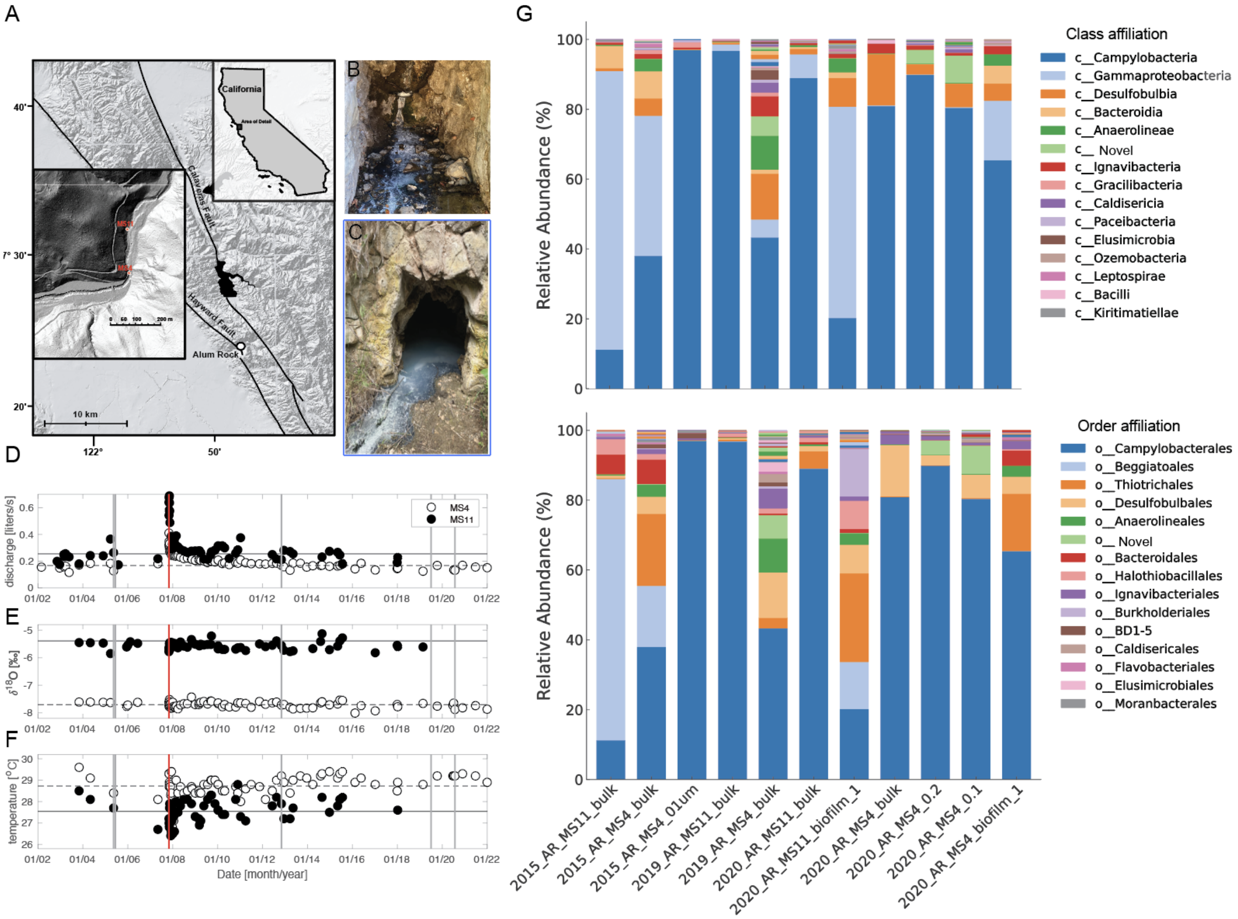


Figure 1.1 A) Shaded relief map showing the location of Alum Rock springs, CA, USA. Insets show the location of Alum Rock and of the MS4 and MS11 springs. Photographs of B) MS4 and C) MS11 biofilms. Thin white streamers (5-10 cm) are mostly found attached to the surfaces of rocks. Hydrogeological properties D) Discharge E) δ18O, and F) Temperature are steady over periods greater than a decade, except following large regional earthquakes. A discharge increases in late 2007 followed a magnitude 5.6 earthquake with an epicenter 4 km from the springs (vertical red line), neither δ18O nor the temperature changed indicating that fluid sources did not change. The horizontal lines show averages of plotted quantities over the entire sampling period, except discharge for which the average excludes the first two years after the earthquake. Vertical grey lines show dates of biofilm and planktonic sampling. G) Microbial community composition at the class and order levels, respectively, highlighting the top 15 most abundant groups in each category. Each bar represents a sample collected from different biotopes (bulk/biofilm, 0.1 µm filter) in the MS4 and MS11 springs over several years (2015, 2019, and 2020). The stacked bar plots illustrate the relative abundance of each microbial group, with each color corresponding to a different group, from top to bottom in decreasing order of overall abundance across all samples.

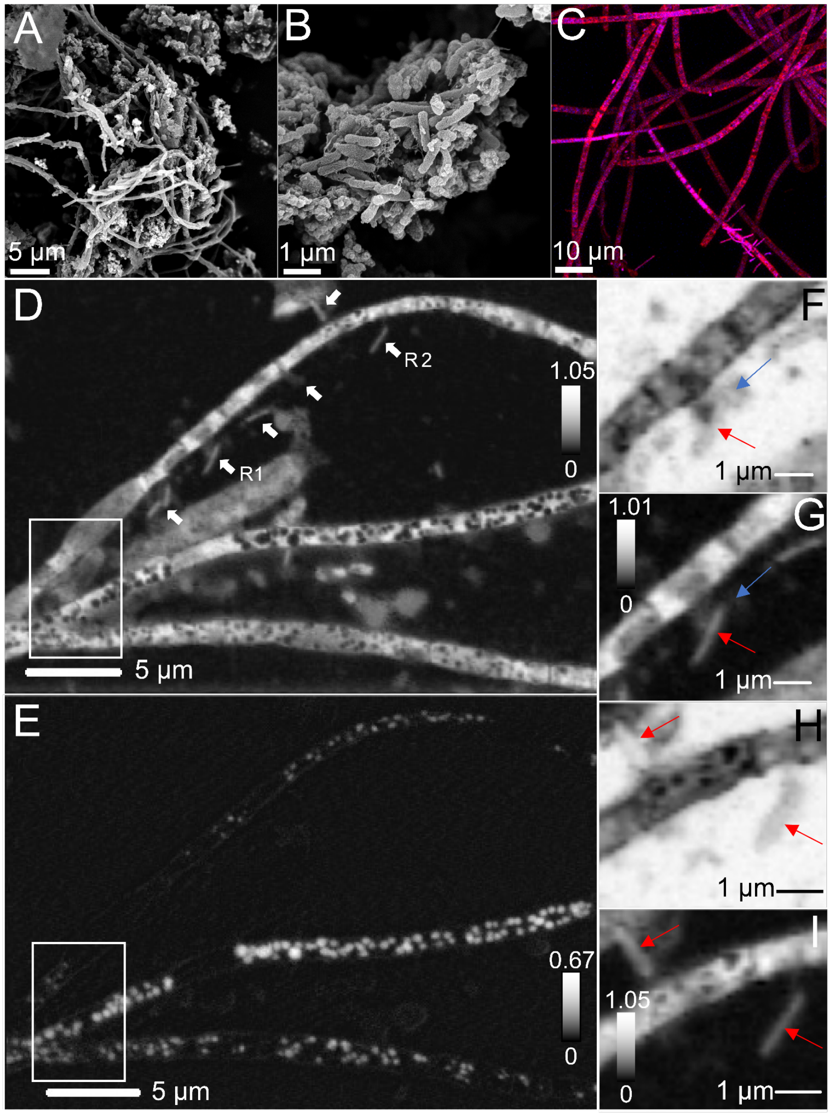


Figure 1.2 Microscopic characterization of MS4 biofilms. A-B) Scanning electron microscopy of filamentous bacteria and associated cells, small cells are pointed by arrows C) Confocal fluorescence microscopy of cells treated with SYTOX (blue) for nucleic acid and F-64 (red) for membrane. Scanning transmission x-ray microscopy: D) Carbon map of filaments and associated cells (white arrows). E) Corresponding distribution map of S^0 evidencing spherical elemental sulfur granules within the compartments of the filaments**.** The top, middle and bottom filament widths are 1.23 ± 0.5 µm, 1.01 ± 0.2 µm and 1.33 ± 0.3 µm respectively. F) An ultra-small cell ~480 nm long, ~270 nm wide, (blue arrow) in contact with an apparently episymbiotic cell 1.86 ± 0.1 µm long, ~360 nm wide (red arrow), imaged at 280 eV (region R1, panel D) and corresponding G) Carbon map. H) Two apparently episymbiotic cells (red arrows) connected to filaments, imaged at 280 eV (region R2, panel D) and corresponding I) Carbon map. The intensity scales correspond to optical density.

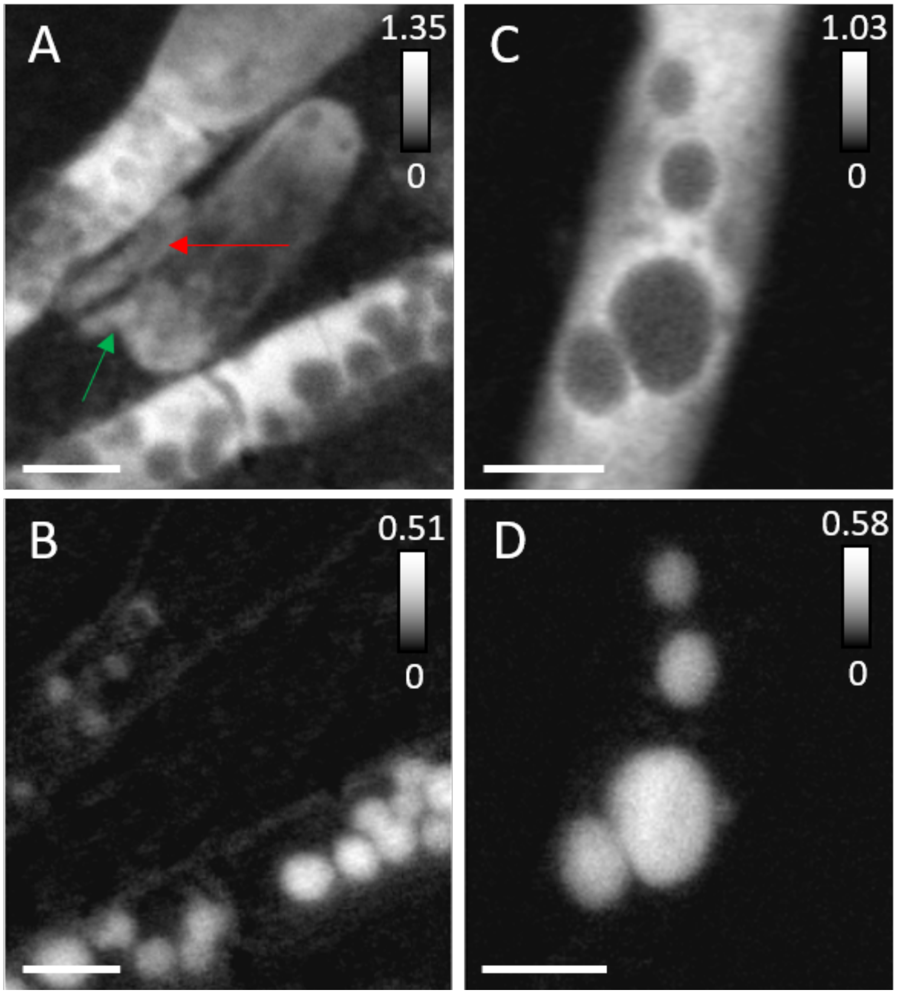


Figure 1.3 Scanning Transmission X-ray Microscopy of MS4 and MS11 biofilms. A) Protein map and corresponding B) Distribution map of S^0 in MS4 biofilms (in white boxed area of Figure 1.2D). Cells that are 908 ± 32 nm long, 370 ± 30 nm wide (red arrow), 687 ± 34 nm long, 244 ± 33 nm wide (green arrow), seen in close contact with filaments. C) Protein map and corresponding D) Distribution map of S^0 in MS11 biofilms, showing the presence of sulfur granules (up to 1.08 ± 0.12 μm in diameter) in a small area of a long filament. Sulfur $L_{2,3}$ edge spectra of the granules can be found in Figure S2. The intensity scale corresponds to the optical density. Scale bars: 1μm.

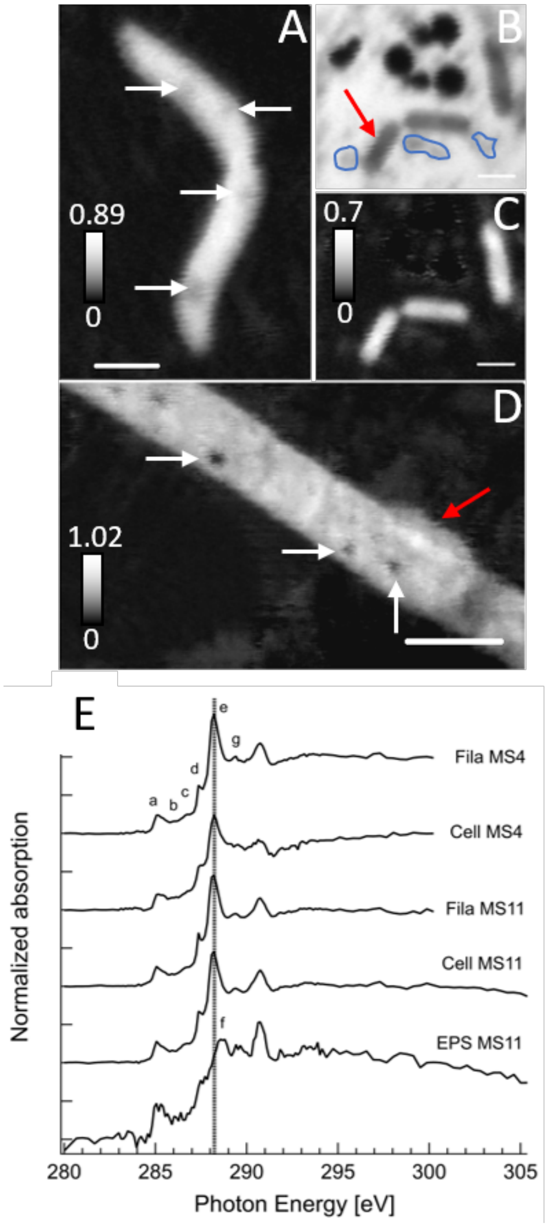


Figure 1.4 Scanning transmission x-ray spectromicroscopy at 87 Kelvin of frozen-hydrated MS4 and MS11 biofilms. A) Protein map of an MS11 small filament (700 nm ±120 nm in diameter), S^0 granules are pointed by white arrows. B) Extracellular S^0 granules (~300 to 850 nm in diameter) and MS11 cells imaged at 288.2 eV (amide group in proteins) and corresponding C) protein map. D) Protein map of an MS4 filament (1.46 ± 0.16 µm in diameter) with a surface-attached cell (1.9 ± 0.21 µm, red arrow). E) Carbon K-edge NEXAFS spectra of filamentous bacteria (S^0 granule-free areas) exhibiting a major peak at 288.2 eV and other peaks mainly associated with nucleic acids (see Table S2). Spectra of MS11 and MS4 cells (red arrow) exhibit a main peak at 288.2 eV (peptide bond) and associated extracellular polymeric substances (EPS, circled in blue) show a main peak at 288.7 eV (carboxyl groups in acidic polysaccharides). See Table S2 for further details. Dashed line is at 288.2 eV. The intensity scale corresponds to the optical density. Scale bars: 1µm (A-C) and 2 µm (D).

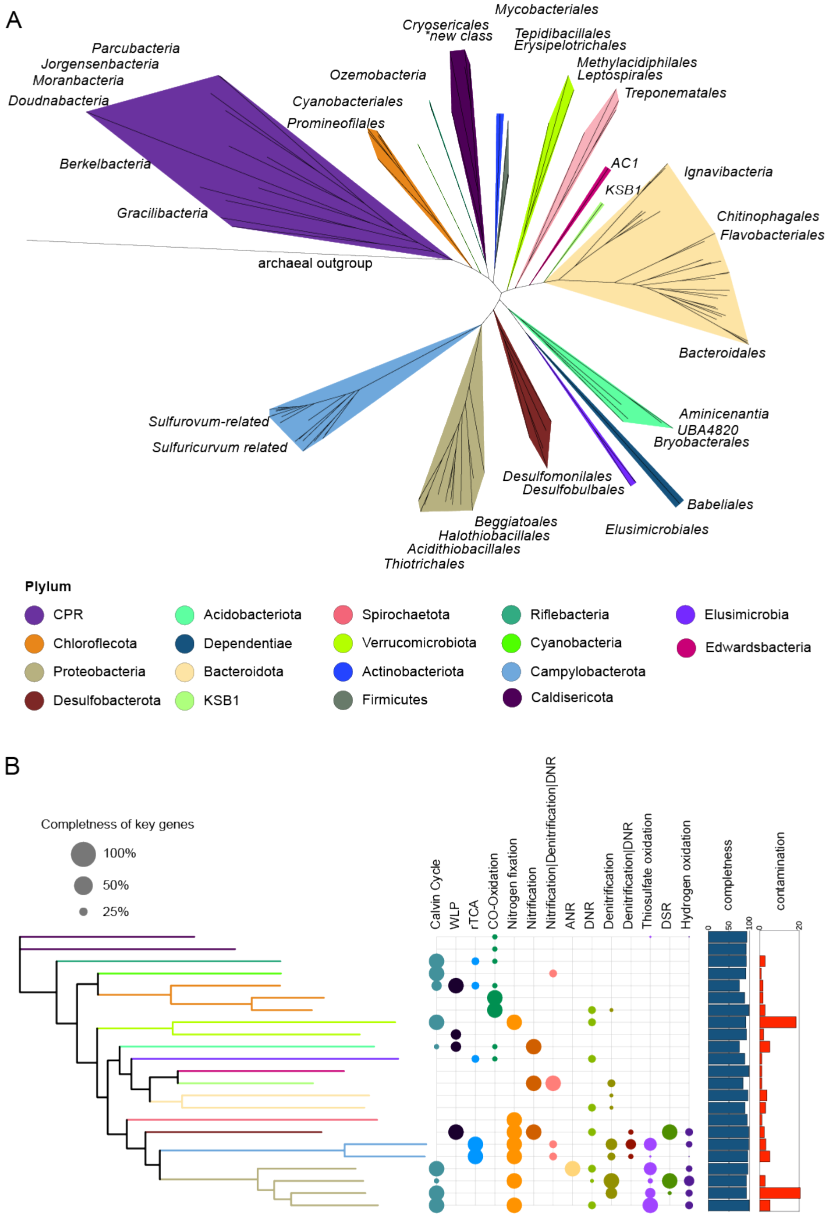


Figure 1.5 Phylogenetic analysis and metabolism of bacteria represented by MAGs from the MS4 and MS11 sites. A) The tree is based on 16 concatenated ribosomal proteins (rpL2, 3, 4, 5, 6, 14, 15, 16, 18, 22, 24 and rpS3, 8, 10, 17, 19) generated using iQ-TREE. An archaeon, *Thermoccocus alcaliphilu*s, was used as the outgroup. B) The metabolic capacities for generalized biogeochemical pathways in Alum Rock genomes are represented by colored circles. A pathway is present if the core KEGG orthologs encoding that pathway are identified in each genome. Abbreviations are as follows: WLP, Wood–Ljungdahl pathway, rTCA, reductive tricarboxylic acid cycle; ANR, Assimilatory nitrate reduction; DNRA, dissimilatory nitrate reduction to ammonia; Thiosulfate oxidation by SOX complex; DSR, Dissimilatory sulfate reduction; Hydrogen oxidation, [NiFe] hydrogenases and NAD-reducing hydrogenase.

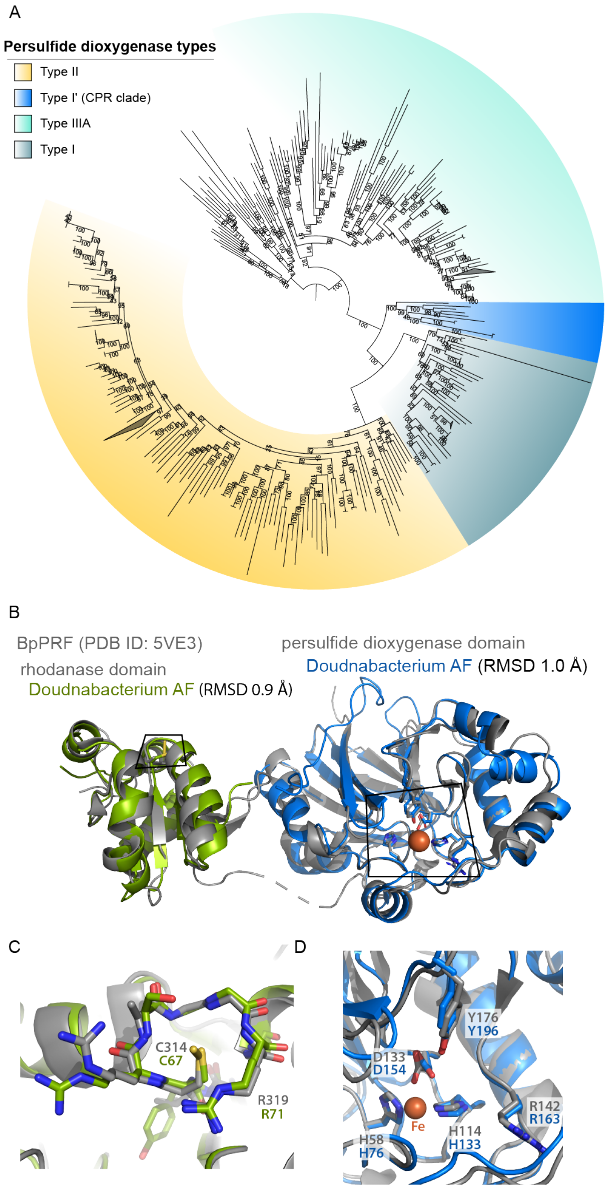


Figure 1.6 Novel Persulfide dioxygenase within CPR Bacteria. A) Phylogenetic analyses of persulfide dioxygenase proteins from the Alum Rock genomic bins. The blue monophyletic clade shows the persulfide dioxygenase found in CPR bacteria from sulfur-rich environments. B) AlphaFold models of Doudnabacterium putative rhodonase (green) and persulfide dioxygenase (blue) aligned with the corresponding domains of the characterized natural fusion protein BpRF (PDB ID: 5VE3). C) and D) Zoomed views of the active sites of the aligned structures reveal a strong coincidence of the key residues.

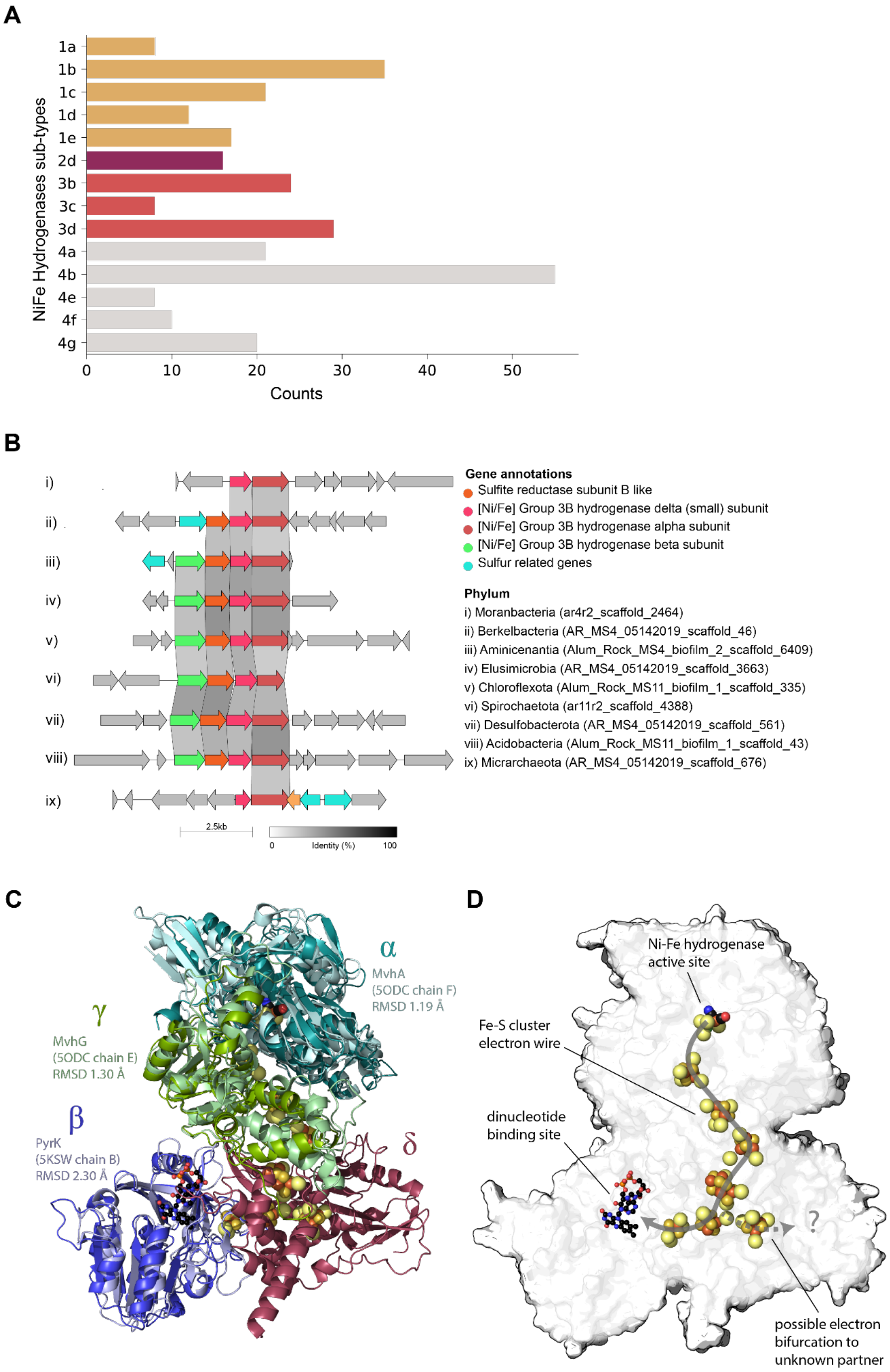


Figure 1.7 Hydrogenases distribution in Alum Rock genomes and structural insights of Group 3b [NiFe]-hydrogenase complex. A) Total distribution of hydrogenases from the Alum Rock spring. B) Genomic organization of novel Group 3b [NiFe]-hydrogenases from different organisms present in the springs. C and D) Alphafold multimeric model for the Berkelbacterium putative Group 3b [NiFe]-hydrogenase complex with the closest known structural matches aligned to each protein.

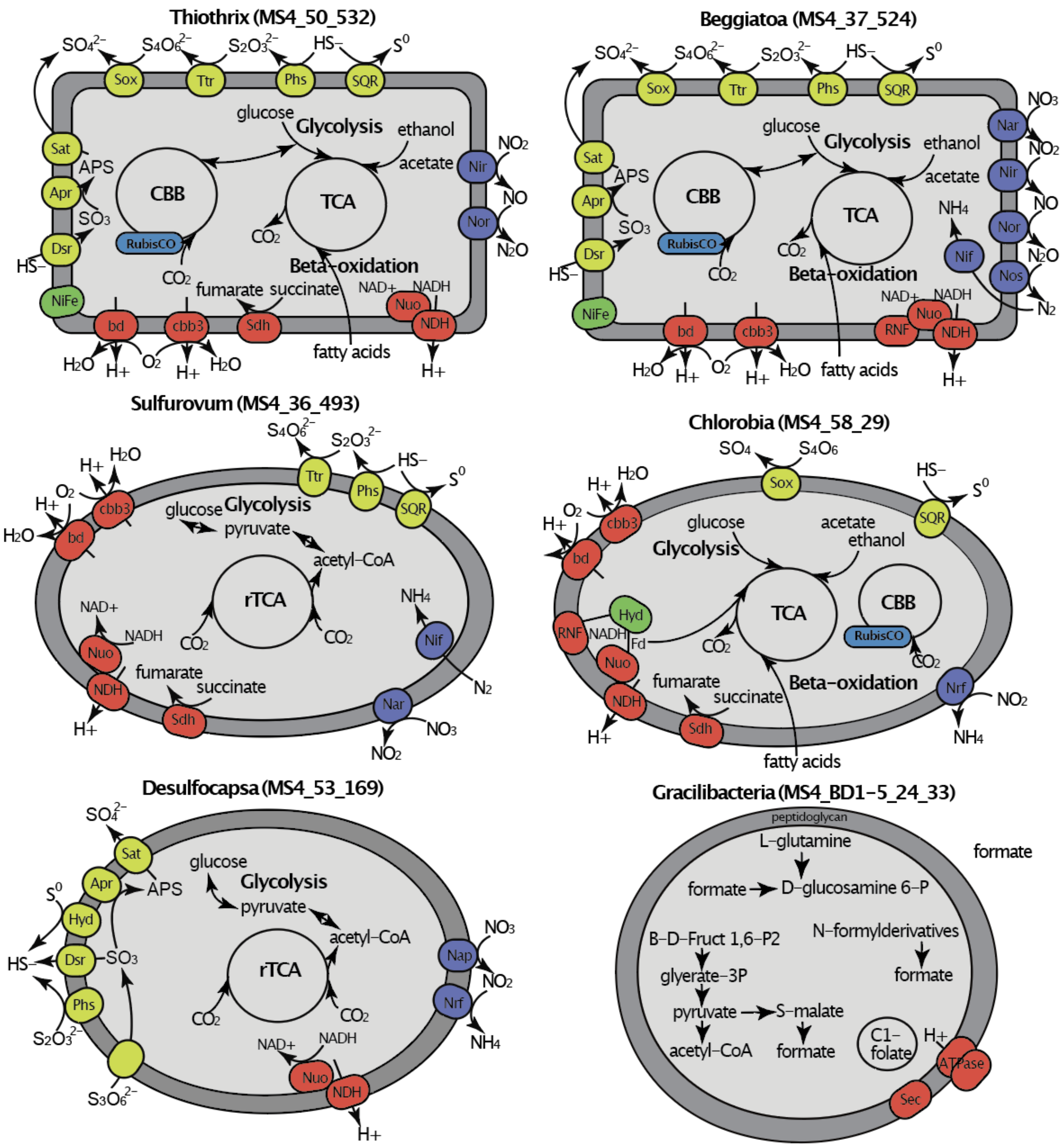


Figure 1.8 Metabolic capacities of several bacteria at Alum Rock spring MS4. Overview of metabolic pathways in six bacterial taxa from Alum Rock spring MS4. This schematic representation details the central metabolic processes, including glycolysis, the tricarboxylic acid (TCA) cycle, and the Calvin-Benson-Bassham (CBB) cycle, across different bacterial taxa. Key enzymes and pathways for the transformation of sulfur, nitrogen, and hydrogen are highlighted, with the presence of specific marker genes such as Sox (sulfur oxidation), Nir (nitrite reduction), and Nif (nitrogen fixation) indicated by colored circles.

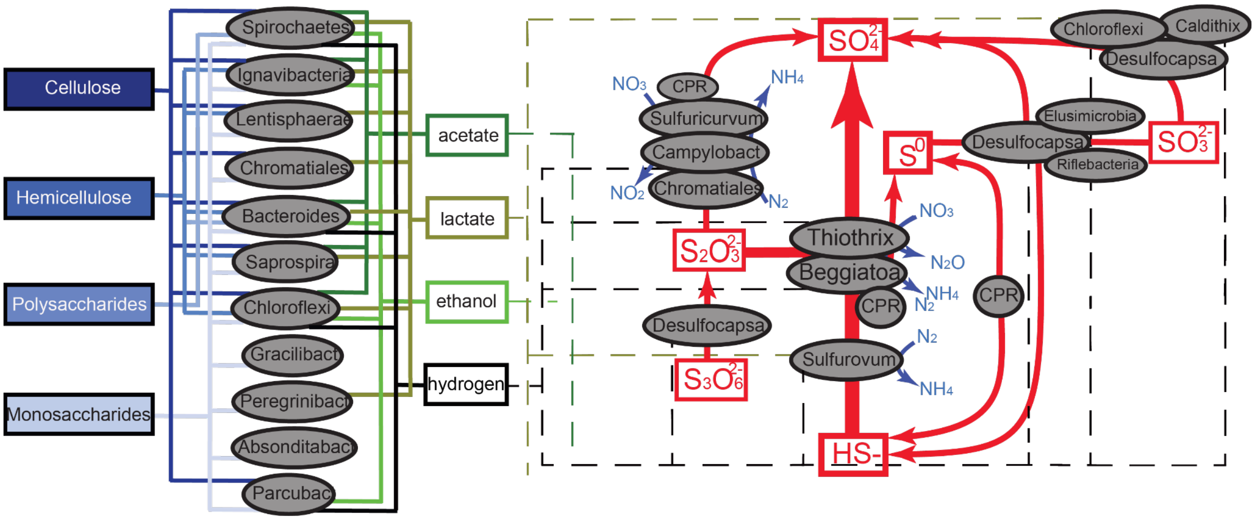
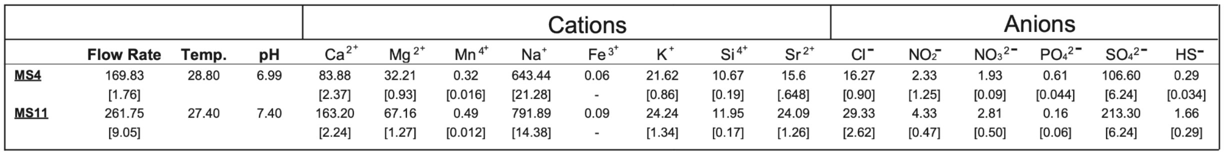


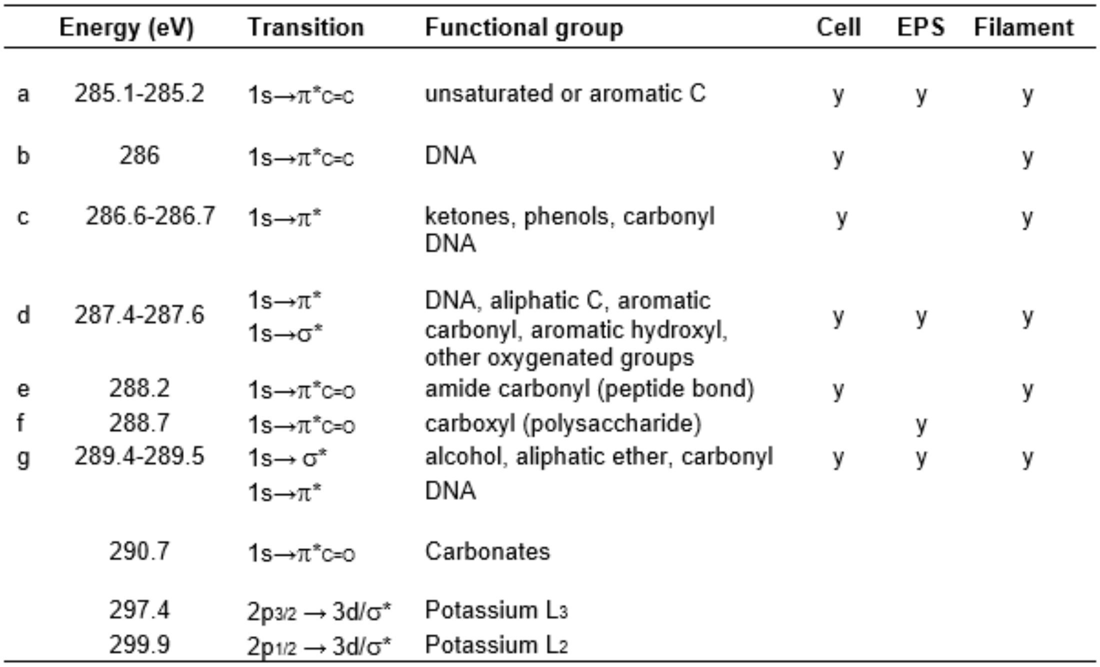
Figure 1.9 Inference of partitioning of carbon, sulfur and nitrogen cycling in the Alum Rock springs. Based on the gene content of genomes reconstructed from the springs. Arrows indicate metabolic capacities reconstructed from metagenomes recovered from MS4 and MS11 springs. The dashed lines represent potential electron donors for anaerobic respiration processes.

1.6 Tables



				Cations								Anions					
	Flow Rate	**Temp.**	**pH**	Ca^{2+}	Mg^{2+}	Mn^{4+}	Na^{+}	Fe^{3+}	K^{+}	Si^{4+}	Sr^{2+}	Cl^{-}	NO_2^{-}	NO_3^{2-}	PO_4^{2-}	SO_4^{2-}	HS^{-}
MS4	169.83	28.80	6.99	83.88	32.21	0.32	643.44	0.06	21.62	10.67	15.6	16.27	2.33	1.93	0.61	106.60	0.29
	[1.76]			[2.37]	[0.93]	[0.016]	[21.28]	-	[0.86]	[0.19]	[.648]	[0.90]	[1.25]	[0.09]	[0.044]	[6.24]	[0.034]
MS11	261.75	27.40	7.40	163.20	67.16	0.49	791.89	0.09	24.24	11.95	24.09	29.33	4.33	2.81	0.16	213.30	1.66
	[9.05]			[2.24]	[1.27]	[0.012]	[14.38]	-	[1.34]	[0.17]	[1.26]	[2.62]	[0.47]	[0.50]	[0.06]	[6.24]	[0.29]

Supplementary Table S1 Geochemical parameters of the two Alum Rock springs in June 2005. Flow rates given in mL/s, temperature in °C, and concentrations in mg/L**.**



	Energy (eV)	Transition	Functional group	Cell	EPS	Filament
a	285.1-285.2	$1s\rightarrow\pi^*_{C=C}$	unsaturated or aromatic C	y	y	y
b	286	$1s\rightarrow\pi^*_{C=C}$	DNA	y		y
c	286.6-286.7	$1s\rightarrow\pi^*$	ketones, phenols, carbonyl DNA	y		y
d	287.4-287.6	$1s\rightarrow\pi^*$ $1s\rightarrow\sigma^*$	DNA, aliphatic C, aromatic carbonyl, aromatic hydroxyl, other oxygenated groups	y	y	y
e	288.2	$1s\rightarrow\pi^*_{C=O}$	amide carbonyl (peptide bond)	y		y
f	288.7	$1s\rightarrow\pi^*_{C=O}$	carboxyl (polysaccharide)		y	
g	289.4-289.5	$1s\rightarrow\sigma^*$ $1s\rightarrow\pi^*$	alcohol, aliphatic ether, carbonyl DNA	y	y	y
	290.7	$1s\rightarrow\pi^*_{C=O}$	Carbonates			
	297.4	$2p_{3/2}\rightarrow 3d/\sigma^*$	Potassium L_3			
	299.9	$2p_{1/2}\rightarrow 3d/\sigma^*$	Potassium L_2			

Supplementary Table S2 Major carbon functional groups present in MS4 and MS11 biofilms, the peaks were assigned according to prior work (Ade et al., 1992; Hitchcock et al., 2002; Kaznacheyev et al., 2002; Samuel et al., 2006; Yabuta et al., 2014)**.**

1.7 References

Ade, H., Zhang, X., Cameron, S., Costello, C., Kirz, J., & Williams, S. (1992). Chemical contrast in X-ray microscopy and spatially resolved XANES spectroscopy of organic specimens. *Science*, *258*(5084), 972–975. https://doi.org/10.1126/science.1439809

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., Thomas, B. C., Singh, A., Wilkins, M. J., Karaoz, U., Brodie, E. L., Williams, K. H., Hubbard, S. S., & Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 13219. https://doi.org/10.1038/ncomms13219

Anantharaman, K., Hausmann, B., Jungbluth, S. P., Kantor, R. S., Lavy, A., Warren, L. A., Rappé, M. S., Pester, M., Loy, A., Thomas, B. C., & Banfield, J. F. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *The ISME Journal*, *12*(7), 1715–1728. https://doi.org/10.1038/s41396-018-0078-0

Beetz, T., & Jacobsen, C. (2003). Soft X-ray radiation-damage studies in PMMA using a cryo-STXM. *Journal of Synchrotron Radiation*, *10*(Pt 3), 280–283. https://doi.org/10.1107/s0909049503003261

Benzerara, K., Yoon, T. H., Tyliszczak, T., Constantz, B., Spormann, A. M., & Brown, G. E. (2004). Scanning transmission X-ray microscopy study of microbial calcification. *Geobiology*, *2*(4), 249–259. https://doi.org/10.1111/j.1472-4677.2004.00039.x

Berney, M., Greening, C., Conrad, R., Jacobs, W. R., Jr, & Cook, G. M. (2014). An obligately aerobic soil bacterium activates fermentative hydrogen production to survive reductive stress during hypoxia. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(31), 11479–11484. https://doi.org/10.1073/pnas.1407034111

Boese, J., Osanna, A., Jacobsen, C., & Kirz, J. (1997). Carbon edge XANES spectroscopy of amino acids and peptides. *Journal of Electron Spectroscopy and Related Phenomena*, *85*(1), 9–15. https://doi.org/10.1016/S0368-2048(97)00032-7

Bor, B., Bedree, J. K., Shi, W., McLean, J. S., & He, X. (2019). Saccharibacteria (TM7) in the Human Oral Microbiome. *Journal of Dental Research*, *98*(5), 500–509. https://doi.org/10.1177/0022034519831671

Bor, B., Collins, A. J., Murugkar, P. P., Balasubramanian, S., To, T. T., Hendrickson, E. L., Bedree, J. K., Bidlack, F. B., Johnston, C. D., Shi, W., McLean, J. S., He, X., & Dewhirst, F. E. (2020). Insights Obtained by Culturing Saccharibacteria With Their Bacterial Hosts. *Journal of Dental Research*, *99*(6), 685–694. https://doi.org/10.1177/0022034520905792

Brandes, J. A., Wirick, S., & Jacobsen, C. (2010). Carbon K-edge spectra of carbonate minerals. *Journal of Synchrotron Radiation*, *17*(5), 676–682. https://doi.org/10.1107/S0909049510020029

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211. https://doi.org/10.1038/nature14486

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* , *25*(15), 1972–

1973. https://doi.org/10.1093/bioinformatics/btp348

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., & Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews. Microbiology*, *16*(10), 629–645. https://doi.org/10.1038/s41579-018-0076-2

Chapelle, F. H., O'Neill, K., Bradley, P. M., Methé, B. A., Ciufo, S. A., Knobel, L. L., & Lovley, D. R. (2002). A hydrogen-based subsurface microbial community dominated by methanogens. *Nature*, *415*(6869), 312–315. https://doi.org/10.1038/415312a

Chaudhari, N. M., Overholt, W. A., Figueroa-Gonzalez, P. A., Taubert, M., Bornemann, T. L. V., Probst, A. J., Hölzer, M., Marz, M., & Küsel, K. (2021). The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. *Environmental Microbiome*, *16*(1), 24. https://doi.org/10.1186/s40793-021-00395-w

Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, *20*(8), 1203–1212. https://doi.org/10.1038/s41592-023-01940-w

Comolli, L. R., & Downing, K. H. (2005). Dose tolerance at helium and nitrogen temperatures for whole cell electron tomography. *Journal of Structural Biology*, *152*(3), 149–156. https://doi.org/10.1016/j.jsb.2005.08.004

Cross, K. L., Campbell, J. H., Balachandran, M., Campbell, A. G., Cooper, C. J., Griffen, A., Heaton, M., Joshi, S., Klingeman, D., Leys, E., Yang, Z., Parks, J. M., & Podar, M. (2019). Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nature Biotechnology*, *37*(11), 1314–1321. https://doi.org/10.1038/s41587-019-0260-6

Danczak, R. E., Johnston, M. D., Kenah, C., Slattery, M., Wrighton, K. C., & Wilkins, M. J. (2017). Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome*, *5*(1), 112. https://doi.org/10.1186/s40168-017-0331-1

Ehrlich, H. L., & Newman, D. K. (2008). *Geomicrobiology, Fifth Edition*. Taylor & Francis. https://doi.org/10.1201/9780849379079

Engel, A. S., Porter, M. L., Stern, L. A., Quinlan, S., & Bennett, P. C. (2004). Bacterial diversity and ecosystem function of filamentous microbial mats from aphotic (cave) sulfidic springs dominated by chemolithoautotrophic "Epsilonproteobacteria." *FEMS Microbiology Ecology*, *51*(1), 31–53. https://doi.org/10.1016/j.femsec.2004.07.004

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2022). Protein complex prediction with AlphaFold-Multimer. In *bioRxiv* (p. 2021.10.04.463034). https://doi.org/10.1101/2021.10.04.463034

Fakra, S. C., Luef, B., Castelle, C. J., Mullin, S. W., Williams, K. H., Marcus, M. A., Schichnes, D., & Banfield, J. F. (2018). Correlative Cryogenic Spectromicroscopy to Investigate Selenium Bioreduction Products. *Environmental Science & Technology*, *52*(2), 503–512. https://doi.org/10.1021/acs.est.5b01409

Finster, K., Liesack, W., & Thamdrup, B. (1998). Elemental sulfur and thiosulfate disproportionation by Desulfocapsa sulfoexigens sp. nov., a new anaerobic bacterium isolated from marine surface sediment. *Applied and Environmental Microbiology*, *64*(1), 119–125. https://doi.org/10.1128/AEM.64.1.119-125.1998

Finster, K. W., Kjeldsen, K. U., Kube, M., Reinhardt, R., Mussmann, M., Amann, R., & Schreiber, L. (2013). Complete genome sequence of Desulfocapsa sulfexigens, a marine deltaproteobacterium specialized in disproportionating inorganic sulfur compounds. *Standards in Genomic Sciences*, *8*(1), 58–68. https://doi.org/10.4056/sigs.3777412

Friedrich, C. G., Bardischewsky, F., Rother, D., Quentmeier, A., & Fischer, J. (2005). Prokaryotic sulfur oxidation. *Current Opinion in Microbiology*, *8*(3), 253–259. https://doi.org/10.1016/j.mib.2005.04.005

Frigaard, N.-U., & Dahl, C. (2009). Sulfur metabolism in phototrophic sulfur bacteria. *Advances in Microbial Physiology*, *54*, 103–200. https://doi.org/10.1016/S0065-2911(08)00002-7

Fritsch, J., Scheerer, P., Frielingsdorf, S., Kroschinsky, S., Friedrich, B., Lenz, O., & Spahn, C. M. T. (2011). The crystal structure of an oxygen-tolerant hydrogenase uncovers a novel iron-sulphur centre. *Nature*, *479*(7372), 249–252. https://doi.org/10.1038/nature10505

Gilchrist, C. L. M., & Chooi, Y.-H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab007

Grote, J., Schott, T., Bruckner, C. G., Glöckner, F. O., Jost, G., Teeling, H., Labrenz, M., & Jürgens, K. (2012). Genome and physiology of a model Epsilonproteobacterium responsible for sulfide detoxification in marine oxygen depletion zones. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(2), 506–510. https://doi.org/10.1073/pnas.1111262109

Hahn, C. R., Farag, I. F., Murphy, C. L., Podar, M., Elshahed, M. S., & Youssef, N. H. (2022). Microbial Diversity and Sulfur Cycling in an Early Earth Analogue: From Ancient Novelty to Modern Commonality. *mBio*, *13*(2), e0001622. https://doi.org/10.1128/mbio.00016-22

Hamilton, T. L., Jones, D. S., Schaperdoth, I., & Macalady, J. L. (2014). Metagenomic insights into S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem. *Frontiers in Microbiology*, *5*, 756. https://doi.org/10.3389/fmicb.2014.00756

Handley, K. M., Bartels, D., O'Loughlin, E. J., Williams, K. H., Trimble, W. L., Skinner, K., Gilbert, J. A., Desai, N., Glass, E. M., Paczian, T., Wilke, A., Antonopoulos, D., Kemner, K. M., & Meyer, F. (2014). The complete genome sequence for putative H_2- and S-oxidizer Candidatus Sulfuricurvum sp., assembled de novo from an aquifer-derived metagenome. *Environmental Microbiology*, *16*(11), 3443–3462. https://doi.org/10.1111/1462-2920.12453

He, C., Keren, R., Whittaker, M. L., Farag, I. F., Doudna, J. A., Cate, J. H. D., & Banfield, J. F. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nature Microbiology*, *6*(3), 354–365. https://doi.org/10.1038/s41564-020-00840-5

He, X., McLean, J. S., Edlund, A., Yooseph, S., Hall, A. P., Liu, S.-Y., Dorrestein, P. C., Esquenazi, E., Hunter, R. C., Cheng, G., Nelson, K. E., Lux, R., & Shi, W. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(1), 244–249. https://doi.org/10.1073/pnas.1419038112

Hinck, S., Neu, T. R., Lavik, G., Mussmann, M., de Beer, D., & Jonkers, H. M. (2007). Physiological adaptation of a nitrate-storing Beggiatoa sp. to diel cycling in a phototrophic hypersaline mat. *Applied and Environmental Microbiology*, *73*(21), 7013–7022. https://doi.org/10.1128/AEM.00548-07

Hitchcock, A. P., Morin, C., Heng, Y. M., Cornelius, R. M., & Brash, J. L. (2002). Towards practical soft X-ray spectromicroscopy of biomaterials. *Journal of Biomaterials Science. Polymer Edition*, *13*(8), 919–937. https://doi.org/10.1163/156856202320401960

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*, 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Inagaki, F., Takai, K., Nealson, K. H., & Horikoshi, K. (2004). Sulfurovum lithotrophicum gen. nov., sp. nov., a novel sulfur-oxidizing chemolithoautotroph within the epsilon-Proteobacteria isolated from Okinawa Trough hydrothermal sediments. *International Journal of Systematic and Evolutionary Microbiology*, *54*(Pt 5), 1477–1482. https://doi.org/10.1099/ijs.0.03042-0

Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., & Banfield, J. F. (2020). The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology*, *18*(1), 69. https://doi.org/10.1186/s12915-020-00804-5

Jaffe, A. L., Konno, M., Kawasaki, Y., Kataoka, C., Béjà, O., Kandori, H., Inoue, K., & Banfield, J. F. (2022). Saccharibacteria harness light energy using type-1 rhodopsins that may rely on retinal sourced from microbial hosts. *The ISME Journal*, *16*(8), 2056–2059. https://doi.org/10.1038/s41396-022-01231-w

Jones, D. S., Albrecht, H. L., Dawson, K. S., Schaperdoth, I., Freeman, K. H., Pi, Y., Pearson, A., & Macalady, J. L. (2012). Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. *The ISME Journal*, *6*(1), 158–170. https://doi.org/10.1038/ismej.2011.75

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalanetra, K. M., Huston, S. L., & Nelson, D. C. (2004). Novel, attached, sulfur-oxidizing bacteria at shallow hydrothermal vents possess vacuoles not involved in respiratory nitrate accumulation. *Applied and Environmental Microbiology*, *70*(12), 7487–7496. https://doi.org/10.1128/AEM.70.12.7487-7496.2004

Kang, D. D., 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*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kaznacheyev, K., Osanna, A., Jacobsen, C., Plashkevych, O., Vahtras, O., Ågren, Carravetta, V., & Hitchcock, A. P. (2002). Innershell Absorption Spectroscopy of Amino Acids. *The Journal of Physical Chemistry. A*, *106*(13), 3153–3168. https://doi.org/10.1021/jp013385w

Kilcoyne, A. L. D., Tyliszczak, T., Steele, W. F., Fakra, S., Hitchcock, P., Franck, K., Anderson, E., Harteneck, B., Rightor, E. G., Mitchell, G. E., Hitchcock, A. P., Yang, L., Warwick, T., & Ade, H. (2003). Interferometer-controlled scanning transmission X-ray microscopes at the Advanced Light Source. *Journal of Synchrotron Radiation*, *10*(Pt 2), 125–136. https://doi.org/10.1107/s0909049502017739

Kirz, J., Jacobsen, C., & Howells, M. (1995). Soft X-ray microscopes and their biological applications. *Quarterly Reviews of Biophysics*, *28*(1), 33–130. https://doi.org/10.1017/s0033583500003139

Kuroda, K., Yamamoto, K., Nakai, R., Hirakata, Y., Kubota, K., Nobu, M. K., & Narihiro, T. (2022). Symbiosis between Candidatus Patescibacteria and Archaea Discovered in Wastewater-Treating Bioreactors. *mBio*, e0171122. https://doi.org/10.1128/mbio.01711-22

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lapidus, A., Nolan, M., Lucas, S., Glavina Del Rio, T., Tice, H., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Pagani, I., Ivanova, N., Huntemann, M., Mavromatis, K., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., … Woyke, T. (2011). Genome sequence of the filamentous, gliding Thiothrix nivea neotype strain (JP2(T)). *Standards in Genomic Sciences*, *5*(3), 398–406. https://doi.org/10.4056/sigs.2344929

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H.-Y. N., Birarda, G., Thomas, B. C., Singh, A., Williams, K. H., Siegerist, C. E., Tringe, S. G., Downing, K. H., Comolli, L. R., & Banfield, J. F. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature Communications*, *6*, 6372. https://doi.org/10.1038/ncomms7372

Macalady, J. L., Dattagupta, S., Schaperdoth, I., Jones, D. S., Druschel, G. K., & Eastman, D. (2008). Niche differentiation among sulfur-oxidizing bacterial populations in cave waters. *The ISME Journal*, *2*(6), 590–601. https://doi.org/10.1038/ismej.2008.25

Macalady, J. L., Lyon, E. H., Koffman, B., Albertson, L. K., Meyer, K., Galdenzi, S., & Mariani, S. (2006). Dominant microbial populations in limestone-corroding stream biofilms, Frasassi cave system, Italy. *Applied and Environmental Microbiology*, *72*(8), 5596–5609. https://doi.org/10.1128/AEM.00715-06

Madrid, V. M., Taylor, G. T., Scranton, M. I., & Chistoserdov, A. Y. (2001). Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. *Applied and Environmental Microbiology*, *67*(4), 1663–1674. https://doi.org/10.1128/AEM.67.4.1663-1674.2001

Ma, K., Schicho, R. N., Kelly, R. M., & Adams, M. W. (1993). Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(11), 5341–5344. https://doi.org/10.1073/pnas.90.11.5341

Marcus, M. A., MacDowell, A. A., Celestre, R., Manceau, A., Miller, T., Padmore, H. A., & Sublett, R. E. (2004). Beamline 10.3.2 at ALS: a hard X-ray microprobe for environmental and materials sciences. *Journal of Synchrotron Radiation*, *11*(Pt 3), 239–247. https://doi.org/10.1107/S0909049504005837

Matheus Carnevali, P. B., Schulz, F., Castelle, C. J., Kantor, R. S., Shih, P. M., Sharon, I., Santini, J. M., Olm, M. R., Amano, Y., Thomas, B. C., Anantharaman, K., Burstein, D., Becraft, E. D., Stepanauskas, R., Woyke, T., & Banfield, J. F. (2019). Hydrogen-based metabolism as an ancestral trait in lineages sibling to the Cyanobacteria. *Nature Communications*, *10*(1), 463. https://doi.org/10.1038/s41467-018-08246-y

Meziti, A., Nikouli, E., Hatt, J. K., Konstantinidis, K. T., & Kormas, K. A. (2021). Time series metagenomic sampling of the Thermopyles, Greece, geothermal springs reveals stable microbial communities dominated by novel sulfur-oxidizing chemoautotrophs. *Environmental Microbiology*, *23*(7), 3710–3726. https://doi.org/10.1111/1462-2920.15373

Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195. https://doi.org/10.1093/molbev/mst024

Moreira, D., Zivanovic, Y., López-Archilla, A. I., Iniesto, M., & López-García, P. (2021). Reductive evolution and unique predatory mode in the CPR bacterium Vampirococcus lugosii. *Nature Communications*, *12*(1), 2454. https://doi.org/10.1038/s41467-021-22762-4

Motl, N., Skiba, M. A., Kabil, O., Smith, J. L., & Banerjee, R. (2017). Structural and biochemical analyses indicate that a bacterial persulfide dioxygenase-rhodanese fusion protein functions in sulfur assimilation. *The Journal of Biological Chemistry*, *292*(34), 14026–14038. https://doi.org/10.1074/jbc.M117.790170

Nelson, D. C., Jørgensen, B. B., & Revsbech, N. P. (1986). Growth Pattern and Yield of a Chemoautotrophic Beggiatoa sp. in Oxygen-Sulfide Microgradients. *Applied and Environmental Microbiology*, *52*(2), 225–233. https://doi.org/10.1128/aem.52.2.225-233.1986

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nielsen, P. H., de Muro, M. A., & Nielsen, J. L. (2000). Studies on the in situ physiology of Thiothrix spp. present in activated sludge. *Environmental Microbiology*, *2*(4), 389–398. https://doi.org/10.1046/j.1462-2920.2000.00120.x

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), 555–560. https://doi.org/10.1038/s41587-020-00777-4

Olson, J. W., & Maier, R. J. (2002). Molecular hydrogen as an energy source for Helicobacter pylori. *Science*, *298*(5599), 1788–1790. https://doi.org/10.1126/science.1077123

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004. https://doi.org/10.1038/nbt.4229

Pedroni, P., Della Volpe, A., Galli, G., Mura, G. M., Pratesi, C., & Grandi, G. (1995). Characterization of the locus encoding the [Ni-Fe] sulfhydrogenase from the archaeon Pyrococcus furiosus: evidence for a relationship to bacterial sulfite reductases. *Microbiology*, *141 (Pt 2)*, 449–458. https://doi.org/10.1099/13500872-141-2-449

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth.

Bioinformatics , *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Raveh-Sadka, T., Thomas, B. C., Singh, A., Firek, B., Brooks, B., Castelle, C. J., Sharon, I., Baker, R., Good, M., Morowitz, M. J., & Banfield, J. F. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *eLife*, *4*. https://doi.org/10.7554/eLife.05477

Ravel, B., & Newville, M. (2005). ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *Journal of Synchrotron Radiation*, *12*(Pt 4), 537–541. https://doi.org/10.1107/S0909049505012719

Rossetti, S., Blackall, L. L., Levantesi, C., Uccelletti, D., & Tandoi, V. (2003). Phylogenetic and physiological characterization of a heterotrophic, chemolithoautotrophic Thiothrix strain isolated from activated sludge. *International Journal of Systematic and Evolutionary Microbiology*, *53*(Pt 5), 1271–1276. https://doi.org/10.1099/ijs.0.02647-0

Rossmassler, K., Hanson, T. E., & Campbell, B. J. (2016). Diverse sulfur metabolisms from two subterranean sulfidic spring systems. *FEMS Microbiology Letters*, *363*(16). https://doi.org/10.1093/femsle/fnw162

Rowland, J. C., Manga, M., & Rose, T. P. (2008). The influence of poorly interconnected fault zone flow paths on spring geochemistry. *Geofluids*, *8*(2), 93–101. https://doi.org/10.1111/j.1468-8123.2008.00208.x

Samuel, N. T., Lee, C.-Y., Gamble, L. J., Fischer, D. A., & Castner, D. G. (2006). NEXAFS characterization of DNA components and molecular-orientation of surface-bound DNA oligomers. *Journal of Electron Spectroscopy and Related Phenomena*, *152*(3), 134–142. https://doi.org/10.1016/j.elspec.2006.04.004

Sarbu, S. M., Kinkle, B. K., Vlasceanu, L., Kane, T. C., & Popa, R. (1994). Microbiological characterization of a sulfide-rich groundwater ecosystem. *Geomicrobiology Journal*, *12*(3), 175–182. https://doi.org/10.1080/01490459409377984

Sarret, G., Connan, J., Kasrai, M., Bancroft, G. M., Charrié-Duhaut, A., Lemoine, S., Adam, P., Albrecht, P., & Eybert-Bérard, L. (1999). Chemical forms of sulfur in geological and archeological asphaltenes from Middle East, France, and Spain determined by sulfur K- and L-edge X-ray absorption near-edge structure spectroscopy. *Geochimica et Cosmochimica Acta*, *63*(22), 3767–3779. https://doi.org/10.1016/S0016-7037(99)00205-7

Sato, T., Atomi, H., & Imanaka, T. (2007). Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science*, *315*(5814), 1003–1006. https://doi.org/10.1126/science.1135999

Schwedt, A., Kreutzmann, A.-C., Polerecky, L., & Schulz-Vogt, H. N. (2011). Sulfur respiration in a marine chemolithoautotrophic beggiatoa strain. *Frontiers in Microbiology*, *2*, 276. https://doi.org/10.3389/fmicb.2011.00276

Sharrar, A. M., Flood, B. E., Bailey, J. V., Jones, D. S., Biddanda, B. A., Ruberg, S. A., Marcus, D. N., & Dick, G. J. (2017). Novel Large Sulfur Bacteria in the Metagenomes of Groundwater-Fed Chemosynthetic Microbial Mats in the Lake Huron Basin. *Frontiers in Microbiology*, *8*, 791. https://doi.org/10.3389/fmicb.2017.00791

Sheik, C. S., Anantharaman, K., Breier, J. A., Sylvan, J. B., Edwards, K. J., & Dick, G. J. (2015). Spatially resolved sampling reveals dynamic microbial communities in rising hydrothermal plumes across a back-arc basin. *The ISME Journal*, *9*(6), 1434–1445. https://doi.org/10.1038/ismej.2014.228

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield,

J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Søndergaard, D., Pedersen, C. N. S., & Greening, C. (2016). HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports*, *6*, 34212. https://doi.org/10.1038/srep34212

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., & Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, *318*(5855), 1449–1452. https://doi.org/10.1126/science.1147112

Stewart-Ornstein, J., Hitchcock, A. P., Hernández Cruz, D., Henklein, P., Overhage, J., Hilpert, K., Hale, J. D., & Hancock, R. E. W. (2007). Using intrinsic X-ray absorption spectral differences to identify and map peptides and proteins. *The Journal of Physical Chemistry. B*, *111*(26), 7691–7699. https://doi.org/10.1021/jp0720993

Stöhr, J. (1992). *NEXAFS Spectroscopy*. Springer Science & Business Media. https://play.google.com/store/books/details?id=N5NBD0393ZYC

Sweerts, J.-P. R. A., Beer, D. D., Nielsen, L. P., Verdouw, H., Van den Heuvel, J. C., Cohen, Y., & Cappenberg, T. E. (1990). Denitrification by sulphur oxidizing Beggiatoa spp. mats on freshwater sediments. *Nature*, *344*(6268), 762–763. https://doi.org/10.1038/344762a0

Takai, K., Campbell, B. J., Cary, S. C., Suzuki, M., Oida, H., Nunoura, T., Hirayama, H., Nakagawa, S., Suzuki, Y., Inagaki, F., & Horikoshi, K. (2005). Enzymatic and genetic characterization of carbon and energy metabolisms by deep-sea hydrothermal chemolithoautotrophic isolates of Epsilonproteobacteria. *Applied and Environmental Microbiology*, *71*(11), 7310–7320. https://doi.org/10.1128/AEM.71.11.7310-7320.2005

Tian, J., Utter, D. R., Cen, L., Dong, P.-T., Shi, W., Bor, B., Qin, M., McLean, J. S., & He, X. (2022). Acquisition of the arginine deiminase system benefits epiparasitic Saccharibacteria and their host bacteria in a mammalian niche environment. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(2). https://doi.org/10.1073/pnas.2114909119

Toner, B. M., Fakra, S. C., Manganini, S. J., Santelli, C. M., Marcus, M. A., Moffett, J. W., Rouxel, O., German, C. R., & Edwards, K. J. (2009). Preservation of iron(II) by carbon-rich matrices in a hydrothermal plume. *Nature Geoscience*, *2*(3), 197–201. https://doi.org/10.1038/ngeo433

Vigneron, A., Cruaud, P., Culley, A. I., Couture, R.-M., Lovejoy, C., & Vincent, W. F. (2021). Genomic evidence for sulfur intermediates as new biogeochemical hubs in a model aquatic microbial ecosystem. *Microbiome*, *9*(1), 46. https://doi.org/10.1186/s40168-021-00999-x

Williams, T. M., & Unz, R. F. (1985). Filamentous sulfur bacteria of activated sludge: characterization of Thiothrix, Beggiatoa, and Eikelboom type 021N strains. *Applied and Environmental Microbiology*, *49*(4), 887–898. https://doi.org/10.1128/aem.49.4.887-898.1985

Wrighton, K. C., Castelle, C. J., Varaljay, V. A., Satagopan, S., Brown, C. T., Wilkins, M. J., Thomas, B. C., Sharon, I., Williams, K. H., Tabita, F. R., & Banfield, J. F. (2016). RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *The ISME Journal*, *10*(11), 2702–2714. https://doi.org/10.1038/ismej.2016.53

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., Wilkins, M. J., Hettich, R. L., Lipton, M. S., Williams, K. H., Long, P. E., & Banfield, J. F. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, *337*(6102), 1661–1665. https://doi.org/10.1126/science.1224041

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* , *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Yabuta, H., Uesugi, M., Naraoka, H., Ito, M., Kilcoyne, A. L. D., Sandford, S. A., Kitajima, F., Mita, H., Takano, Y., Yada, T., Karouji, Y., Ishibashi, Y., Okada, T., & Abe, M. (2014). X-ray absorption near edge structure spectroscopic study of Hayabusa category 3 carbonaceous particles. *Earth, Planets and Space*, *66*(1), 1–8. https://doi.org/10.1186/s40623-014-0156-0

Yakushev, E. V., Pollehne, F., Jost, G., Kuznetsov, I., Schneider, B., & Umlauf, L. (2007). Analysis of the water column oxic/anoxic interface in the Black and Baltic seas with a numerical model. *Marine Chemistry*, *107*(3), 388–410. https://doi.org/10.1016/j.marchem.2007.06.003

Zhang, J., Liu, R., Xi, S., Cai, R., Zhang, X., & Sun, C. (2020). A novel bacterial thiosulfate oxidation pathway provides a new clue about the formation of zero-valent sulfur in deep sea. *The ISME Journal*, *14*(9), 2261–2274. https://doi.org/10.1038/s41396-020-0684-5

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., & Anantharaman, K. (2020). METABOLIC: High-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. In *bioRxiv* (p. 761643). https://doi.org/10.1101/761643

Zubavichus, Y., Shaporenko, A., Korolkov, V., Grunze, M., & Zharnikov, M. (2008). X-ray absorption spectroscopy of the nucleotide bases at the carbon, nitrogen, and oxygen K-edges. *The Journal of Physical Chemistry. B*, *112*(44), 13711–13716. https://doi.org/10.1021/jp802453u

Transitional section

The study of biofilms in mineral sulfide-rich springs has revealed intricate microbial communities that play crucial roles in biogeochemical cycling, particularly in the context of sulfur and hydrogen metabolism. These biofilms, dominated by chemolithotrophic sulfur-oxidizing bacteria and hosting diverse CPR bacteria, provide a unique opportunity to investigate the complex interactions and metabolic adaptations of microorganisms in these environments. Building upon these findings, we found potential novel hydrogenases, many of which come from uncultivable organisms. The lack of experimental validation for these hydrogenases motivated us to start a collaboration with Prof. Chris Greening to investigate microbial hydrogen cycling, a fundamental process that underpins the diversity and functionality of various anoxic ecosystems.

2. Unique minimal and hybrid hydrogenases are active from anaerobic archaea

The following chapter is a modified version with permission of authors of the following pre-print: Chris Greening*, Princess R. Cabotaje*, Luis E. Valentin Alvarado*, Pok Man Leung, Henrik Land, Moritz Senger, Max A. Klamke, Michael Milton, Rachael Lappan, Susan Mullen, Jacob West-Roberts, Jie Mao, Jiangning Song, Marie Schoelmerich, Courtney W. Stairs, Rhys Grinter, Anja Spang, Jillian F. Banfield1 and Gustav Berggren (2023). *In press Cell*. Unique minimal and hybrid hydrogenases are active in anaerobic archaea
*contributed equally to this work

Abstract

Microbial hydrogen (H_2) cycling underpins the diversity and functionality of diverse anoxic ecosystems. Among the three evolutionarily distinct hydrogenase superfamilies responsible, [FeFe]-hydrogenases were thought to be restricted to bacteria and eukaryotes. Here we show that anaerobic archaea encode diverse, active, and ancient lineages of [FeFe]-hydrogenases through combining analysis of existing and new genomes with extensive biochemical experiments. [FeFe]-hydrogenases are encoded by genomes of nine archaeal phyla and expressed by H_2-producing Asgard archaeon cultures. We report a novel ultra-minimal hydrogenase in DPANN archaea that binds the catalytic H-cluster and produces H_2. Moreover, we identified and characterised remarkable hybrid complexes formed through the fusion of [FeFe]- and [NiFe]-hydrogenases in ten other archaeal orders. Phylogenetic analysis and structural modelling suggest a deep evolutionary history of hybrid hydrogenases. These findings reveal new metabolic adaptations of archaea, streamlined H_2 catalysts for biotechnological development, and a surprisingly intertwined evolutionary history between the two major H_2-metabolizing enzymes.

N.B. *All main figures for this manuscript can be found below in section 4.6. All supplementary files (including figures and tables) can be found online with the published bioRxiv preprint.*

2.1 Introduction

Molecular hydrogen (H_2) is heralded as a future green energy carrier. In a biological context, this energy-rich gas already plays a central role in bioenergetics and evolution has driven an elaborate hydrogen economy. Numerous bacteria, archaea, and microbial eukaryotes consume and produce hydrogen gas ($H_2 \rightleftharpoons 2\ H+ + 2\ e-$) using metalloenzymes called hydrogenases (Greening et al. 2016; Schwartz and Friedrich 2006; Peters et al. 2015). Three hydrogenases have independently evolved in microorganisms, namely the [FeFe]-, [NiFe]-, and [Fe]-hydrogenases, which differ in their metal cofactors and catalytic mechanism (Volbeda et al. 1995; Peters et al. 1998; Shima et al. 2008). H_2 serves multiple roles in microbial physiology. Microorganisms spanning all three domains produce H_2 to dispose of electrons during fermentation. Numerous bacteria and archaea also use electrons derived from H_2 oxidation for respiration and carbon fixation (Greening et al. 2016; Vignais and Billoud 2007; Schwartz and Friedrich 2006; Pinske 2019; Leung et al. 2022). More recently, electron-bifurcating hydrogenase complexes have been discovered that are critical for energy conservation in obligate anaerobes (Schut and Adams 2009; Schuchmann and Müller 2012; Wang et al. 2013; Feng et al. 2022; Buckel and Thauer 2013; Schuchmann et al. 2018). Hydrogenases are now recognised as environmentally ubiquitous and taxonomically widespread, encoded in genomes from most bacterial and archaeal phyla, as well as many unicellular eukaryotes (Greening et al. 2016; Peters et al. 2015; Søndergaard et al. 2016). It is increasingly recognized that microbial H_2 metabolism shapes global biogeochemical cycling (Piché-Choquette and Constant 2019; Constant et al. 2009), supports biodiversity of diverse ecosystems (Greening et al. 2022; Morita 1999) and influences health and disease (Benoit et al. 2020; Carbonero et al. 2012). In addition, these efficient enzymes have growing industrial applications in the developing H_2 economy (Cracknell et al. 2008) and serve as an inspiration for the design of synthetic catalyst (Evans et al. 2019; Kleinhaus et al. 2021). H_2 was likely the primordial electron donor, but continues to have a central role in microbiology both as a desirable energy source and diffusible electron sink (Lane et al. 2010; Weiss et al. 2016). Moreover, it is proposed that H_2 exchange between bacteria and archaea underlies eukaryogenesis, as described in various syntrophy hypothese (Martin and Müller 1998; Moreira and Lopez-Garcia 1998; Sousa et al. 2016; Spang et al. 2019; Imachi et al. 2020).

The three hydrogenase classes differ in their physiological roles and taxonomic distribution. [FeFe]-hydrogenases are typically fast-acting, but oxygen-sensitive, and are best known for their roles in obligate anaerobes. These enzymes currently comprise four phylogenetically distinct groups (groups A to D), which can be further subdivided through two different schemes based on domain architecture and genetic organisation (Land et al. 2020; Calusinska et al. 2010; Greening Chris et al. 2016). They include monomeric enzymes that couple ferredoxin oxidation to fermentative H_2 production (group A1) (Peters et al. 1998), trimeric enzymes that reversibly bifurcate electrons from H_2 to NAD+ and ferredoxin (group A3 (Schut and Adams 2009), filamentous complexes with formate dehydrogenase that catalyse H_2-dependent CO2 conversion to formate (group A4) (Dietrich et al. 2022), putative sensory hydrogenases in which the catalytic hydrogenase domain is fused with a PAS domain (group C) (Land et al. 2019), and several functionally undefined groups (e.g. groups B and D) (Greening et al. 2016; Land et al. 2019). Despite this diversity, [FeFe]-hydrogenases are all predicted to rely on the same organometallic cofactor for catalysis, the "H-cluster". To date, these enzymes have been exclusively characterised in anaerobic bacteria and eukaryotes, and appear to be absent in cultured archaea (Greening et al. 2016; Peters et al. 2015; Land et al. 2020). [NiFe]-hydrogenases are extraordinarily structurally and functionally diverse enzymes encoded by bacteria and archaea across all ecosystems. They are presently subdivided into four major groups (groups 1 to 4) and 29 subgroups that each differ in their phylogeny, genetic organisation, and physiological roles (Greening et al. 2016; Vignais and Billoud 2007; Ortiz et al. 2021). The catalytic (large) subunit and electron-relaying iron-sulfur (small) subunit of the [NiFe]-hydrogenase associate with other subunits depending on the subgroup; the different complexes formed can mediate respiration, fermentation, energy-conversion, electron-bifurcation, carbon fixation, and H_2 sensing processes 1. In contrast, [Fe]-hydrogenases are a much narrower lineage that contribute to archaeal methanogenesis (Shima et al. 2008; Schleucher et al. 1994). The three hydrogenase classes are phylogenetically unrelated, despite having some similar structural features, and are not thought to genetically or structurally associate. [NiFe]-hydrogenases are predicted to have been present in the last universal common ancestor (LUCA), whereas [FeFe]-hydrogenases are proposed to have evolved later in fermentative bacteria (Greening et al. 2016; Weiss et al. 2016).

In the eight years since hydrogenase distribution has been comprehensively surveyed 1, there has been a massive expansion of known microbial diversity primarily due to genome-resolved metagenomics (i.e. the recovery of microbial genomes from mixed community samples) (Tyson et al. 2004; Rinke et al. 2013; Castelle and Banfield 2018; Spang et al. 2017). This expansion has been particularly pronounced in the domain archaea (Spang et al. 2017). Notable developments include the discovery of the Asgard superphylum from which eukaryotes are predicted to have evolved (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017; Rambo et al. 2022; Rodrigues-Oliveira et al. 2023; Eme et al. 2023), the rapid expansion of the DPANN superphylum that include obligate symbionts (Rinke et al. 2013; Castelle et al. 2015; Castelle et al. 2018; Dombrowski et al. 2019), and the discovery of new lineages capable of anaerobic alkane metabolism (Laso-Pérez et al. 2016; Chen et al. 2019; Evans et al. 2019; Dong et al. 2019; Borrel et al. 2019). Most of these archaea are uncultivated, necessitating culture-independent approaches to characterise them. Many novel lineages appear to be capable of H_2 metabolism. For example, the recently cultured Asgard archaeon ('Candidatus Prometheoarchaeum syntrophicum') fermentatively generates H_2, seemingly through the activity of an electron-bifurcating/confurcating [NiFe]-hydrogenase, which is consumed by its syntrophic methanogenic partner (Imachi et al. 2020). The H_2-metabolising enzymes reported in these newly discovered archaea fall into a range of established and novel [NiFe]-hydrogenase subgroups (Sousa et al. 2016; Spang et al. 2019; Castelle et al. 2018; Dombrowski et al. 2019). Several genomic studies have also suggested that [FeFe]-hydrogenases may be encoded by uncultivated DPANN archaea (Greening et al. 2016; Castelle et al. 2018; Huang et al. 2021). However, this remains debatable given archaea seemingly lack the three maturation enzymes (HydEFG) required to synthesize the biologically unique catalytic H-cluster of the [FeFe]-hydrogenase (Mulder et al. 2010), and to date there is no evidence for archaeal [FeFe]-hydrogenase activity.

Here we systematically analysed the diversity of [FeFe]-hydrogenases in the domain archaea by searching all publicly available species-representative genomes and MAGs, as well as multiple new MAGs. We relied on two innovations to validate these findings. First, we used AlphaFold2-based structural modelling to test whether conserved gene clusters encode novel hydrogenase complexes (Jumper et al. 2021; Evans et al. 2022). Critically, we then combined heterologous enzyme production with artificial maturation to confirm whether archaeal hydrogenases were catalytically active and displayed H-cluster spectroscopic signals (Berggren et al. 2013; Khanna et al. 2017; Land et al. 2019). Through this integrated approach, we demonstrate that archaea harbour active ultraminimal [FeFe]-hydrogenases and identify novel complexes with distinct sequences, structures, and probable functions to previously described enzymes, including novel hybrid complexes with [NiFe]-hydrogenases. These findings revise our understanding of the distribution and evolution of microbial H_2 metabolism, and have broad biological, chemical, and biotechnological ramifications.

2.2 Materials and Methods

Genome sources

In this study, 130 archaeal genomes encoding [FeFe]-hydrogenases were analysed. 40 genomes are novel metagenome-assembled genomes retrieved from our unpublished datasets at ggKbase (https://ggkbase-help.berkeley.edu/). genomes were retrieved from the Genome Taxonomy Database (GTDB) R06-RS202 (Parks et al. 2018) following a search of all 2,339 archaeal species representative genomes. GTDB was chosen as the data source as it offers a standardised taxonomy, a manageable search space due to pre-clustered sequences, and genomes pre-annotated with gene predictions (Parks et al. 2022) In addition, 13 other Asgardarchaeota genomes were retrieved from PATRIC 102. Full details of the genomes analysed are provided in Table S1. All 130 genomes were examined for completeness and contamination with CheckM (Parks et al. 2015) and manually curated through taxonomic profiling in ggKbase. Initial taxonomic classification of genome bins was performed with GTDB-Tk-R202 and subsequently confirmed by constructing phylogenetic trees. All archaeal genomes analysed were dereplicated at 95% average nucleotide identity using dRep v.3.3 (Olm et al. 2017).

Metagenomic assembly and binning

The 40 newly reported genomes were retrieved from a range of anoxic ecosystems. Most MAGs came from previously reported study sites, namely Guaymas Basin hydrothermal vents (7 MAGs; Gulf of California, Mexico) (Castelle et al. 2021), Lac Lavin freshwater lake (6 MAGs; central France) (Jaffe et al. 2023), Crystal Geyser (6 MAGs; Utah, USA) (Probst et al. 2017), Chinese hot springs (5 MAGs; 2 from Yunnan Province and 3 from Tibetan Plateau, China) (De Anda et al. 2021), an aquifer (3 MAGs; Napa County, California, USA) (He et al. 2021), Manure lagoon (2 MAGs, California, USA) (He et al. 2021), an aquifer adjacent to the Colorado River (3 MAGs; Rifle, Colorado, USA) (Anantharaman et al. 2016), a wetland soil (2 MAGs; Napa County, California, USA) (Al-Shayeb et al. 2022), Alum Rock mineral spring (1 MAG; San Jose, California) (Valentin-Alvarado et al. 2024), Zodletone Spring (1 MAG; Anadarko, Oklahoma, USA), Azore Islands hot springs (1 MAGs, Portugal) (Méheust et al. 2020), a borehole (1 MAG, Muzunami, Japan). Sampling, DNA extraction, sequencing library preparation, and sequencing methods were previously described. Briefly, metagenomic sequencing reads assembled using IDBA-UD (Peng et al. 2012). Contigs larger than 2.5 kb were retained, and sequencing reads from all samples were mapped against each resulting assembly utilizing Bowtie2 (Langmead and Salzberg 2012). Differential coverage profiles, filtered with a 95% read identity threshold, were then used for genome binning using a suite of binning tools (MetaBAT2 (Kang et al. 2019), VAMB (Nissen et al. 2021), MaxBin2 (Wu et al. 2016, Abawaca (https://github.com/CK7/abawaca), with the final bin choice determined by DAS Tool (Sieber et al. 2018). Additionally, 1 MAG was obtained from samples from Corona Mine drainage at the Oat Hill Mine, Napa County, California, USA; in this case, water was filtered through 2.5 and 0.1 μm filters sequentially, DNA was extracted from each filter separately, and two runs of Illumina paired-end sequencing were performed at 150 bp and 250 bp read lengths. 1 MAG were also obtained from hyporheic zone water sampled from beneath the riverbed of the East River, Gunnison County, Colorado, USA; in this case, water was filtered through a 0.1 μm filter, DNA was extracted from the filter, and Illumina sequencing was performed with a read length of 150 bp. For the East River and Oat Hill Mine samples, genome data were assembled using MetaSPAdes v3.15.5 (Nurk et al. 2017) with default k-mer values. Draft genomes reconstructed using the MetaBAT2 (Kang et al. 2019), VAMB (Nissen et al. 2021), MaxBin2 (Wu et al. 2016), and via ggKbase manual binning tools and the best genome selected using DAS Tool (Sieber et al. 2018).

Identification of archaeal [FeFe]-hydrogenase genes

The process of identifying [FeFe]-hydrogenase enzymes in archaea involved matching a "training" profile of known [FeFe]-hydrogenases against a database of "candidate" protein sequences. Representative archaeal genomes from GTDB R06-RS202 were used as the data source for candidate hydrogenases (Parks et al. 2018). The training data was the catalytic domains of the [FeFe]-hydrogenases identified in the HydDB (Søndergaard et al. 2016). The analysis pipeline was written using the Nextflow pipeline framework, which allowed for the analysis to be run reproducibly in containers, which were executed in parallel across nodes of the computing cluster (Di Tommaso et al. 2017). The first process in the pipeline performed a multiple sequence alignment of training sequences, which was then used to build a hidden Markov model (HMM) using hmmer3 (Eddy 2011) that modelled the primary sequence of the [FeFe]-hydrogenase catalytic domain. Candidate protein sequences with length greater than 100,000 amino acids were first excluded from the analysis. Each protein sequence from each representative genome was then matched against the HMM to obtain a bit score, which represented the similarity of the candidate sequence to the known enzyme profile. Preliminary matches with a positive bit score were retained and then filtered by the presence of the CxxxC motif required to ligate the [FeFe]-hydrogenase catalytic centre. The matches passing this filter were then aggregated into a report, along with their taxonomy and bit score. Following this, the matches were manually inspected through a combination of Conserved Domain Database (CDD) annotations (Lu et al. 2020) and phylogenetic analysis to derive a bit score cutoff. A final cutoff of 15.9 was chosen to include all true positives and exclude all other sequences.

Analysis of [FeFe]-hydrogenase domain architecture

[FeFe]-hydrogenases are often multidomain proteins comprising a conserved H_2-activating domain named the H-cluster and various accessory domains at the N- and C-terminals involved in electron transfer or H_2 sensing (Greening et al. 2016; Land et al. 2020). To locate the positions of the H-cluster, all retrieved archaeal [FeFe]-hydrogenase sequences were aligned against the trimmed reference [FeFe]-hydrogenase H-cluster sequences from HydDB 18 using Clustal Omega v1.2.2 (default setting) (Sievers and Higgins 2018). N- and C-terminal regions outside of the H-cluster were extracted and annotated against CDD v3.19 (Lu et al. 2020) using rpsblast (-evalue 0.01 -max_hsps 1 -max_target_seqs 10) in BLAST+ v2.9.0 (Camacho et al. 2009). The N-terminal fusion of the small subunit of [NiFe]-hydrogenase with certain archaeal [FeFe]-hydrogenases was confirmed by searching against Pfam protein family database v34.0 (Mistry et al. 2021) and protein structural modelling as detailed below. The N- and C-terminal sequences of [FeFe]-hydrogenases were further searched for the presence of signature iron-sulfur cluster motifs or specific pattern of cysteine residues that are potentially involved in electron transfer: plant ferredoxin-like [2Fe2S] (Cx10-11Cx2Cx11-15C); bacterial ferredoxin-like 2[4Fe4S] (Cx2Cx2Cx3Cx23-32Cx2Cx2Cx3C); (Cys)3His-ligated [4Fe4S] (Hx2-3Cx2Cx5C) (Winkler et al. 2013). Based on the domain organization and the classification scheme by (Land et al. 2020), archaeal [FeFe]-hydrogenases were assigned into different subclasses. Domain annotation and subclass classification results of the [FeFe]-hydrogenases are summarized in Table S2.

Analysis of [FeFe]-hydrogenase genetic organization

To characterize the genetic context and potential interacting proteins of archaeal [FeFe]-hydrogenases, up to 10 genes upstream and downstream of the catalytic subunits were retrieved.

These neighboring genes were annotated against CDD v3.19 (Lu et al. 2020) using rpsblast (-evalue 0.01 -max_hsps 1 -max_target_seqs 5) in BLAST+ v2.9.0 (Camacho et al. 2009), the Pfam protein family database v34.0 (Mistry et al. 2021) using PfamScan v1.6 (default setting)(Mistry et al. 2007), and the NCBI RefSeq protein database release 202 (Pruitt et al. 2007) using DIAMOND v0.9.31 blastp algorithm (--max-hsps 1 --max-target-seqs 1) (Buchfink et al. 2015). Protein subcellular localization and the presence of internal helices of the gene were predicted using PSORTb v3.0.3 (--archaea) (Yu et al. 2010). The subgroup lineage of the flanking gene that encodes large subunit of [NiFe]-hydrogenase was classified using HydDB (Søndergaard et al. 2016). To facilitate curation of the annotation results and identification of conserved neighboring genes, all flanking genes were clustered at an identity threshold of 30% and a minimum coverage of 80% using MMseqs2 (--min-seq-id 0.3 -c 0.8 --cov-mode 1 --cluster-mode 2 -s 7.5) (Steinegger and Söding 2018). The R package gggenes v0.4.1 (https://github.com/wilkox/gggenes) was used to construct gene arrangement diagrams. All sequences, annotations, and genetic arrangements are summarized in Table S2.

Identification of archaeal [FeFe]-hydrogenase maturases

To probe the maturation pathway and evolution of [FeFe]-hydrogenases in Archaea, we performed a genomic survey of the three conserved [FeFe]-hydrogenase maturases (HydE, HydF, HydG) in all representative archaeal species in GTDB R06-RS202 (Parks et al. 2022). We first compiled a comprehensive database of known [FeFe]-hydrogenase maturase sequences from bacteria and eukaryotes and expanded the dataset based on BLAST searches against the NCBI non-redundant protein database (Nov 2021) (Pruitt et al. 2007). New and authentic divergent hits including from archaea were included in the final database, which was then used to screen for the presence of [FeFe]-hydrogenase maturases in archaeal genomes. To enable the discovery of phylogenetically novel maturases, a relaxed default setting of the DIAMOND v0.9.31 BLASTp algorithm was first applied (Buchfink et al. 2015). False positive hits were filtered by further searching against CDD v3.19 (Lu et al. 2020) for the presence of rSAM_HydE (TIGR03956), rSAM_HydG (TIGR03955), and GTP_HydF (TIGR03918) domains. The [FeFe]-hydrogenase maturase databases and hits from the archaeal genomes are provided in Table S3.

Transcriptomic analysis

‘Ca. Lokiarchaeum ossiferum’ B35 enrichments were grown in MLM medium 53 at 20°C. The cultures were sampled every 14 days and DNA was extracted using the NucleoSpin Soil DNA extraction kit (Macherey-Nagel) according to the manufacturer’s instructions. Growth was monitored by qPCR assays using primers LkF (5′- ATC GAT AGG GGC CGT GAG AG) and LkR (5′- CCC GAC CAC TTG AAG AGC TG) targeting lokiarchaeal 16S rRNA genes as previously described (Rodrigues-Oliveira et al. 2023). When ‘Ca. L. ossiferum’ B35 cells reached mid to late exponential growth (~8.0 × 106 16S rRNA gene copies/mL), 35 mL of enrichment culture was centrifuged at 20,000 × g for 30 minutes, 4°C. The resulting pellet was used for RNA extraction, which was performed using the mirVana miRNA isolation kit (Invitrogen), according to the manufacturer’s instructions. To remove any leftover genomic DNA, the samples were incubated with TURBO DNase (Invitrogen) at 37°C for 1 h. The absence of DNA was confirmed through lokiarchaeal 16S rRNA gene PCR assays. Metatranscriptomic analysis was performed using the NovaSeq PE150 platform through the services of the company Novogene (United Kingdom). The RNA-seq data were analyzed with Trimmomatic v0.36 (Bolger et al. 2014) to remove adaptor sequences and filtering low-quality reads using the following parameters:

LEADING:3 TRAILING:3 MINLEN:36. High quality reads from both replicates (s1 and s2) were mapped to the genome of 'Ca. L. ossiferum' B35 using bowtie2 v2.5.1 (Langmead and Salzberg 2012). Transcript expression was normalized into transcripts per million (TPM) using StringTie v2.2.1 (Pertea et al. 2015).

Proteomic analysis

Pellets of exponentially phase 'Ca. L. ossiferum' enrichments were suspended in 100 μL iST lysis buffer (Preomics). To lyse cells, the suspensions were heated (95°C, 1,000 rpm, 10 min), sonicated (water bath, 3 mins), homogenized (Beatbox (Preomics), 10 mins, high setting), and further heated (95°C, 1,000 rpm, 10 min). Samples were then centrifuged at 10,000 × g for 1 min and 50 μL supernatants were digested using the iST kit following the vendor's protocol (Preomics). Peptides were separated on a Vanquish Neo nano-flow chromatography system (Thermo-Fisher), using a pre-column for sample loading (Acclaim PepMap C18, 2 cm × 0.1 mm, 5 μm, Thermo-Fisher), and a C18 analytical column (Acclaim PepMap C18, 50 cm × 0.75 mm, 2 μm, Thermo-Fisher), applying a segmented linear gradient from 2% to 35% and finally 80% solvent B (80 % acetonitrile, 0.1 % formic acid; solvent A 0.1 % formic acid) at a flow rate of 230 nL min-1 over 120 min. Eluting peptides were analyzed on an Exploris 480 Orbitrap mass spectrometer (Thermo Fisher), which was coupled to the column with a FAIMS pro ion-source (Thermo-Fisher) using coated emitter tips (PepSep, MSWil). The mass spectrometer was operated in DIA mode with the FAIMS CV set to -45, the survey scans were obtained in a mass range of 350-1200 m/z, at a resolution of 60k at 200 m/z and a normalized AGC target at 300%. 60 MSMS spectra with variable isolation width between 7 and 191 m/z covering 349.5-1200.5 m/z range, including 1 m/z windows overlap, were acquired in the HCD cell at 30% collision energy at a normalized AGC target of 1000% and a resolution of 30k. The max. injection time was set to auto. Raw data were processed using Spectronaut software v.17.6 (https://biognosys.com/software/spectronaut/) with the DirectDIA+ workflow. A custom database that included the Ca. L. ossiferum B35 proteins and the most common contaminants was used. The searches were performed with full trypsin specificity and a maximum of 2 missed cleavages at a protein and peptide spectrum match false discovery rate of 1%. Carbamidomethylation of cysteine residues were set as fixed, oxidation of methionine and N-terminal acetylation as variable modifications. The cross-run normalization was turned off and all other settings were left as default. Computational analysis was performed using Python and the in-house developed Python library MsReport v.0.0.14 120. Protein intensity less than 1000 was set to missing to remove the low-quality quantification values. LFQ protein intensities reported by SpectroNaut were log2-transformed and normalized across samples using the ModeNormalizer from MsReport. The missing normalized LFQ intensity values were imputed by drawing random values from a normal distribution after filtering out contaminants, proteins with less than 2 peptides and less than 1 quantified values in at least one group.

AlphaFold2 structural modelling

[FeFe]-hydrogenase models were generated using AlphaFold multimer v2.1.1 implemented on the Monash University MASSIVE M3 computing cluster (Jumper et al. 2021; Tunyasuvunakool et al. 2021). The amino acid sequences for HydA from each of the [FeFe]-hydrogenase groups (A1, A3, B, E, and F) shown in Table S5 and below were modelled both alone and with the sequences of putative complex partners present in the hydrogenase gene cluster. Modelling with putative complex partners was performed iteratively, with output models assessed to determine if a credible complex was generated. Predicted complexes were then modelled with higher stoichiometries, where possible given limitations of GPU RAM (~2,500 amino acids), to predict larger order structures. Models produced were validated based on confidence scores (pLDDT) with only regions with a confidence score of >85 utilised for analysis. Where complexes were predicted subunit interfaces were inspected manually for surface complementarity and the absence of clashing atoms. Interfaces were also analysed for stability using the program QT-PISA (Krissinel 2015; Krissinel and Henrick 2007), with only interfaces predicted to be stable utilised for analysis. To assign cofactors to [FeFe]-hydrogenase models generated by AlphaFold the closest homologous structures or domains were identified by searching the PDB database using NCBI BLAST or the DALI server (Holm and Laakso 2016; Camacho et al. 2009). The homologous structures were aligned with the AlphaFold models and cofactors were added in corresponding positions to that of the experimental structures, providing all conserved coordinating residues were present. Cofactor position was then manually adjusted to optimize coordination and to minimise clashes. The PDB IDs for the structures were used to model cofactors for the archaeal hydrogenases are:

Group A1 – UBA95 sp002499405 (Micrarchaeota), 'Ca. Iainarchaeum andersonii': HydA = 6GM0 Chain A

Group A3 – DSAL01 sp011380095 (Altarchaeota): HydA = 6GM0 Chain A; HydB (NuoF-like) = 7E5Z chain B; HydC (NuoE-like) = 7E5Z chain A

Group B – 'Ca. Prometheoarchaeum syntrophicum': HydA = 6GM0 Chain A ([FeFe]-hydrogenase domain), 7BKD Chain A (ferredoxin domain); HydC (NuoE-like) = 8E9G Chain E

Group E – CABMGN01 sp902385635 (Nanoarchaeota), 'Ca. Forterrea multitransposorum': HydA = 6GM0 Chain A

Group F – UBA147 sp002496385 (Thermoplasmatota): HydA = 6GM0 chain A ([FeFe]-hydrogenase domain), 5ODC chain E ([NiFe]-hydrogenase small domain); HyhL = 5ODC chain L; HydD (NuoG-like) = 7T2R chain A, 5ODC chain E; HydB (NuoF-like) = 7E5Z chain B, 7MFM chain I; HydC (NuoE-like) = 7E5Z chain A

Protein expression and characterization

All chemicals used during the protein production and characterization were purchased from VWR and used as received unless otherwise stated. The genes encoding group A, B, E, and F [FeFe]-hydrogenases (Table S1) were codon optimized for expression in *E. coli*, synthesized, and cloned in pET-11a(+) by Genscript using restriction sites NdeI and BamHI following codon optimization for expression in *E. coli*. Expression constructs were re-transformed in chemically competent E. coli BL21(DE3) cells to express the apo-forms of the hydrogenases lacking the diiron subsite of the H-cluster. Starter cultures were grown overnight in 5 mL LB medium containing 100 µg mL-1 ampicillin at 37°C. These cultures were subsequently used to inoculate 80 mL of M9 medium (22 mM Na_2HPO_4, 22 mM KH_2PO_4, 85 mM NaCl, 18 mM NH_4Cl, 0.2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.4% (v/v) glucose) containing 100 µg mL-1 ampicillin. Cultures were grown at 37°C and 150 rpm until an optical density (OD_{600}) of approximately 0.4 to 0.6 was reached. Protein expression was induced by the addition of 0.1 mM $FeSO_4$ and 1 mM IPTG. Induced cultures were incubated at 20°C and 150 rpm for approximately 16 h. Cells were thereafter harvested by centrifugation at 4,930 × g for 10 mins at 4°C. All subsequent operations were carried out under anaerobic conditions to prevent hydrogenase inactivation by atmospheric oxygen in an MBRAUN glovebox ([O_2] < 5 ppm). The cell pellet was resuspended in a 0.5 mL lysis buffer (30 mM Tris-HC pH 8.0, 0.2 % (v/v) Triton X-100, 0.6 mg mL-1 lysozyme, 0.1 mg mL-1 DNase, 0.1 mg mL-

1 RNase). Cell lysis was performed by three cycles of freezing/thawing in liquid N_2, and the supernatant was recovered by centrifugation (29,080 × g, 10 mins, 4°C).

Hydrogenase semisynthetic maturation

The subsite [2Fe]H subsite mimic, $(Et_4N)_2[Fe_2(SCH_2NHCH_2S)(CO)_4(CN)_2]$ ([2Fe]adt), was synthesized in accordance to literature protocols with minor modifications and verified by FTIR spectroscopy (Li and Rauchfuss 2002; Zaffaroni et al. 2012; Schmidt et al. 1999; Lyon et al. 1999). Incorporation of cofactor was performed by addition of 100 µg the [2Fe]adt subsite mimic (final concentration 80 µM) to 380 µL of the supernatant in potassium phosphate buffer (100 mM, pH 6.8) and 1 % (v/v) Triton X-100. The reaction mixture was enclosed in an airtight vial and was anaerobically incubated at 20°C for 1-4 hr. The non-purified lysate containing the [2Fe]adt subsite mimic was mixed with 200 µL of potassium phosphate buffer (100 mM, pH 6.8) with 10 mM methyl viologen and 20 mM sodium dithionite. Reactions were incubated at 37°C up to 120 mins. H_2 production was determined by analyzing the reaction headspace every 15 mins using a PerkinElmer Clarus 500 gas chromatograph (GC) equipped with a thermal conductivity detector (TCD) and a stainless-steel column packed with Molecular Sieve (60/80 mesh). The operational temperatures of the injection port, oven, and detector were 100°C, 80°C, and 100°C, respectively. Argon was used as carrier gas at a flow rate of 35 mL $min-1$. The three biological replicates were run at varying times (1-4 hours) of incubating the cell lysates with the [2Fe]adt subsite mimic. Incubation time was not found to influence the observed H_2 production. Thus, variation in H-cluster formation rates did not appear to have a substantial influence on the outcome of the screening process.

Hydrogenase purification

Expression construct with verified sequence of the group E [FeFe]-hydrogenase from 'Ca. Forterrea multitransposorum' (Fm) was retransformed in chemically competent E. coli Origami™ B(DE3) cells since the strain yielded the highest expression levels from the small-scale expression tests. Starting cultures were grown overnight in 10 mL LB medium containing 100 µg/mL ampicillin and 15 µg/mL kanamycin at 37°C. These cultures were subsequently used to inoculate 1 L of M9 medium (22 mM Na_2HPO_4, 22 mM KH_2PO_2, 85 mM NaCl, 18 mM NH_4Cl, 0.2 mM MgSO4, 0.1 mM $CaCl_2$, 0.4% (v/v) glucose) containing 100 µg/mL ampicillin and 15 µg/mL kanamycin. Cultures were grown at 37°C and 150 rpm until an optical density (OD_{600}) of approximately 0.4 was reached. Protein expression was induced by the addition of 0.1 mM FeSO4 and 1 mM IPTG. Induced cultures were incubated at 20°C and 150 rpm for approximately 16 h. Cells were thereafter harvested by centrifugation in a Beckman Coulter Avanti J-25 centrifuge (4,424 × g, 10 min). All subsequent operations were carried out under anaerobic conditions in the glovebox to prevent hydrogenase inactivation by atmospheric oxygen. The cell pellet was resuspended in 100 mM Tris-HCl pH 8.0, with NaCl (150 mM), $MgCl_2$ (10 mM), lysozyme from chicken egg white (1 mg/mL), DNAse I from bovine pancreas (0.05 mg/mL), RNase A from bovine pancreas (0.05 mg/mL), and a tablet of complete™ EDTA-free protease inhibitor cocktail, and was incubated inside the glovebox for 30 min. Cell lysis was performed by three cycles of freezing/thawing in liquid N_2. Cell debris was removed by centrifugation in a Beckman Coulter Optima L-90K Ultracentrifuge (222,592 × g, 60 min). The supernatant was collected and filtered (0.45 µm syringe filter) before being loaded on a StrepTrapTM XT (Cytiva) affinity column using a BioLogic DuoFlow™ FPLC system (Bio-Rad) and purified according to the manufacturer´s instructions. Products eluted by 50 mM biotin were concentrated using Amicon®Ultra 30 kDa molecular weight cut-off (MWCO) centrifugal filters (Merck Millipore Ltd.). The StrepTrapTM eluates were further separated using size exclusion chromatography via Superdex™ 200 10/300 GL, equilibrated in 100 mM Tris-HCl, 150 mM NaCl pH 8.0, to acquire purified Fm (Figure S9). Samples were stored anaerobically at −80 °C. Coomassie-stained SDS-PAGE was used to assess purity. Protein estimations were performed via Bradford assay using bovine serum albumin as a standard 154. Quantification of Fe-content was performed using a previously reported assay 155 using a commercially available Fe^{2+} standard for AAS TraceCERT (Sigma Aldrich) for the calibration curve.

Activation of purified hydrogenase

To semi-enzymatically reconstitute the iron-sulfur clusters of Fm, a solution of 50 µM apoprotein in 100 mM Tris-HCl, 150 mM NaCl pH 8.0 was incubated with 500 µM dithiothreitol (DTT) under strictly anaerobic conditions for 10 min at room temperature. The iron and sulfur sources were ferrous ammonium sulfate and L-cysteine, respectively, both added in 1.5-fold molar excess to the desired number of Fe-atoms to be added. Reconstitution was initiated by adding a 1% molar equivalent of recombinant cysteine desulfurase (E. coli IscS), slowly releasing sulfide in situ from cysteine. Reaction mixtures were incubated at room temperature up to 2 hours. At the same time, the increase of absorbance around 405 nm was monitored by UV/Vis (Figure S9). The reconstitution process was stopped by running the reaction mixture through a PD-10 column (GE Healthcare), equilibrated in 100 mM Tris-HCl, 150 mM NaCl pH 8.0. To activate Fm with [2Fe]adt, under strictly anaerobic conditions, the reconstituted enzyme (50 µM) was mixed with sodium dithionite (1 mM, 20× excess) in 100 mM phosphate buffer, pH 6.8 and incubated in room temperature for 10 minutes. Cofactor incorporation started with the addition of [2Fe]adt (600 µM, 12× excess), and the reaction mixture was incubated for 1 h. The mixture was loaded onto a PD-10 desalting column (GE Healthcare) equilibrated with 10 mM Tris-HCl pH 8.0. The sample was concentrated using Amicon®Ultra 30 kDa MWCO centrifugal filters (Merck Millipore Ltd.), aliquoted into PCR tubes, and transferred into airtight serum vials (3-5 µL each) before they were flash-frozen in liquid N_2 and stored at -80°C until further use.

Protein film electrochemistry

Protein film electrochemistry experiments were carried out under anaerobic conditions at 20°C and pH 7.0. The three-electrode system was made up of (1) Ag/AgCl (4 M KCl) as reference electrode, (2) rotating disk 5 mm OD pyrolytic graphite edge (PGE) plane (epoxy encapsulated) as working electrode, and (3) graphite rod as the counter electrode. The gas-tight glass cell used featured a water jacket for temperature control and a cell gas inlet/outlet for hydrogen flow control. The buffer used was composed of 5 mM MES, 5 mM CHES, 5 mM HEPES, 5 mM TAPS, 5 mM sodium acetate (NaOAc), with 0.1 M Na_2SO_4 as carrying electrolyte titrated with H_2SO_4 to pH 7.0, and purged with N2 for 3 to 4 hours. The PGE working electrodes were polished with P1200 sandpaper and rinsed with purified water before they were brought into the glovebox. To remove residual O_2 in the PGE electrode, cyclic voltammograms were run at 100 mV/s from -100 to -600 mV (vs. standard hydrogen electrode (SHE)) for 40 scans. The cyclic voltammogram of the blank electrode (no enzyme immobilized) was then recorded at 10 mV/s with the working electrode rotated at 3 krpm. Polycationic polymyxin B sulfate (5 µL of 0.2 mg/mL) was added onto the deaerated PGE surface before adding 5 uL of 5 µM activated enzyme. The mixture was left for 10 min for maximal adsorption before the excess solution was removed by pipet. The cyclic voltammogram of the system with the immobilized enzyme was then recorded at 10 mV/s under

1 atm of H_2. Electrochemical data was acquired using an Eco/Chemie PGSTAT10 and the GPES software (Metrohm/Autolab). Data were analyzed using Origin 8 software. All values are referenced versus SHE. Experiments were conducted on two independent enzyme films, with each film scanned at least three times.

ATR-FTIR spectroscopy

2 µL enzyme solution (110 µM of Fm) in 10 mM Tris buffer (pH 8) was deposited on the ATR crystal. The ATR unit (BioRadII from Harrick) was sealed with a custom build PEEK cell that allowed for gas exchange and illumination mounted in a FTIR spectrometer (Vertex V70v, Bruker) 129. The sample was dried under 100% nitrogen gas and rehydrated with a humidified aerosol (100 mM Tris-HCl, pH 8). Spectra were recorded with 2 cm-1 resolution, a scanner velocity of 80 Hz and averaged of varying number of scans (mostly 1000 Scans). All measurements were performed at ambient conditions (room temperature and pressure, hydrated enzyme films). Photochemical reduction was achieved through a previously established protocol (Lorenzi et al. 2022; Senger et al. 2019). In short, an enzyme film prepared as described above also including Eosin Y (6 mM, 0.5 µL) and triethanolamine (TEOA, 200 mM, 2 µL) was illuminated using a Schott KL2500LCD cold light source. Spectra shown in Figure 3 are representative examples of two technical replicates. Data were analysed using OPUS and Origin 2019 Software.

Genome-wide metabolic annotations

For all archaeal genomes encoding [FeFe]-hydrogenases, genes were predicted using Prodigal v2.6.3 122. Preliminary functional annotations were established and cross-referenced using KEGG HMMs, UniRef100 and UniProt, and collections of metabolic capacities in genome bins were reviewed using ggKbase genome summaries, as previously described 56. Each open reading frame in the archaeal genomes was assigned a KEGG Orthology if the best-scoring KEGG HMM surpassed the bitscore cutoff. For a targeted profiling of the metabolic capacity of the archaeal genomes, we performed a homology-based search against our custom curated database (https://doi.org/10.26180/c.5230745) consisting of metabolic marker genes involved in carbon fixation (RbcL, AcsB, AclB, Mcr, Hbs), alternative electron donors (FdhA, CoxL, CooS, McrA, MmoA, PmoA, IsoA, FCC, Sqr, Sor, SoxB, PsaA, PsbA, ARO), alternative electron acceptors (AsrA, DsrA, NarG, NapA, NirS, NirK, NrfA, NosZ, NorB, Nod, MtrB, OmcB, YgfK, RdhA), respiration (SdhA, FrdA, CoxA, CcoN, CyoA, CydA, AtpA) using DIAMOND v.2.0.11 with filtering cutoffs described previously (Ortiz et al. 2021). Briefly, hits were filtered based on a minimum query cover of 80% and an identity threshold of 50%, except for CoxL, MmoA, and RbcL (all 60%), AtpA, ARO, IsoA, PsbA, and YgfK (all 70%), Hbs (75%), and PsaA (80%). The functional annotations were expanded to include marker genes for fermentation, fatty acid degradation, aromatic compound degradation, carbohydrate metabolism, and sulfur metabolism using curated HMMs from METABOLIC (Zhou et al. 2020) and KofamKOALA v1.3.0 (Aramaki et al. 2020) with default bitscore thresholds. Hits were further inspected and filtered through searching the NCBI CDD database and final annotation results are summarized in Table S6. The final curated metabolic marker genes were visualized by a custom Python script to construct a heatmap that displayed the presence or absence of particular gene markers within the archaeal genomes. Counts of genes by genome were transformed into binary format, where "1" denotes the presence of at least one hit for a specific gene marker, while "0" signifies the absence of hits for the given marker.

Genome-level phylogenetic analysis

To construct the archaeal genome tree, 15 conserved syntenic ribosomal proteins 158 were retrieved, aligned, and concatenated using GOOSOS (https://github.com/jwestrob/GOOSOS) using all archaeal genomes at least 60% complete and less than 5% contaminated (based on CheckM (Parks et al. 2015)). Genomes containing at least 75% of the 15 syntenic proteins were retained. The concatenated ribosomal protein sequences were aligned using MAFFT (Katoh and Standley 2013), followed by trimming with trimAl 126 using the -gt 0.1 option. The final length of the trimmed concatenated protein alignment was 3224 amino acids for 118 genomes. Branch support was obtained using the ultrafast bootstrap method 159 implemented in IQ-TREE v1.6.12 (Nguyen et al. 2015) and the phylogeny was estimated utilizing the following parameters -bb 1000 -m LG+F+G4. All trees were visualized using iTOL v6.3.2 (Letunic and Bork 2021).

Gene-level phylogenetic analysis

The amino acid sequences of [FeFe]-hydrogenases catalytic subunits (HydA) were retrieved from three datasets: the archaeal genomes analysed in this study, all reference sequences from the hydrogenase database (HydDB) (Søndergaard et al. 2016), and additional eukaryotic genomes 89. For all datasets, taxonomy was assigned to each sequence using ETE v3.0.0 111 and CD-HIT v4.6 106 was used to reduce the dataset at the 80% amino acid sequence identity level. The multiple sequences retrieved were aligned using MAFFT v7.304 (settings: --localpair --maxiterate 1000 --reorder) (Katoh and Standley 2013). The resulting alignment was trimmed using trimAl (settings: -gt 0.1) (Capella-Gutiérrez et al. 2009) and manually inspected with Geneious (Kearse et al. 2012) to remove partial sequences (Dataset S1). Using ModelFinder with default parameters implemented in IQ-TREE v1.6.1 114, the best-fit model according to Bayesian information criterion (BIC) was determined to be LG+C20+R+F. Maximum likelihood phylogenetic trees using the LG+F+G4 substitution model were constructed using IQ-TREE v1.6.1 (Nguyen et al. 2015) (setting: -m LG+C20+R+F -nt AUTO -bb 1000 -alrt 1000). To ensure the robustness of the phylogenetic inference, trees were built with 1000 ultrafast bootstraps and 1000 aLRT (Approximate Likelihood Ratio Test) (Guindon et al. 2010) (Dataset S2 & S3). The models were applied to 3,677 amino acid sequences of the catalytic subunit (HydA) of [FeFe]-hydrogenases, including novel hybrid hydrogenases.

2.3 Results and Discussion

Structurally and genetically diverse [FeFe]-hydrogenases are encoded by nine archaeal phyla

We searched for the gene encoding the catalytic subunit of [FeFe]-hydrogenases (HydA) in the 2,339 archaeal species clusters of the Genome Taxonomy Database (GTDB) and our repository of novel archaeal metagenome-assembled genomes. In total, 130 archaeal genomes (90 previously reported, 40 novel) encoded [FeFe]-hydrogenases, spanning nine phyla and 17 classes (Figure 2.1). Except for some Asgard archaea, only one [FeFe]-hydrogenase was present in each genome. The enzymes were verified and classified based on analysis of domain structure (Table S2; Figure S1), genetic organisation (Table S2; Figure S3), maturases (Table S3; Figure 2.1), and primary phylogeny (detailed below). The enzymes fell into six distinct groups, namely the canonical groups A1 (n = 26), A3 (n = 44), and B (n = 12) and the novel groups E (n = 30), F (n = 21), and G (n = 3) (Figure 2.1; Table S2). The archaeal group A1 and group E [FeFe]-hydrogenases are putative fermentative enzymes encoded by three DPANN phyla (Iainarchaeota, Micrarchaeota, Nanoarchaeota). With average sequences of just 363 and 286 residues respectively (after excluding any truncated sequences), these enzymes are much smaller than the most minimal hydrogenase previously characterised (*Chlamydomonas reinhardtii* HydA1; 457 residues 64,70) (Table S2). The most widespread hydrogenase, however, is the electron-bifurcating/confurcating group A3 [FeFe]-hydrogenase. This hydrogenase, together with its partner diaphorase (HydB) and thioredoxin (HydC) subunits, is encoded by at least six DPANN phyla, Thermoplasmatota (class E2), and some Asgard archaea (class Lokiarchaeia) (Figure 2.1).

Among cultured archaea, [FeFe]-hydrogenases are only encoded in the Asgard archaeal enrichment cultures. We detected group B [FeFe]-hydrogenases in the genomes of both 'Candidatus Prometheoarchaeon syntrophicum' (Imachi et al. 2020) (Figure 2.2C; Table S1) and our recently reported culture 'Candidatus Lokiarchaeum ossiferum' (Rodrigues-Oliveira et al. 2023) (Table S4). While 'Ca. P. syntrophicum' has previously been inferred to fermentatively produce H_2, this activity was assumed to originate from its group 3c [NiFe]-hydrogenase and its [FeFe]-hydrogenase was misannotated as an F420H_2-dependent dehydrogenase subunit; based on previously reported transcriptomes (Imachi et al. 2020), the [FeFe]-hydrogenase is expressed at similarly high levels (309 RPKM) to the [NiFe]-hydrogenase (333 RPKM), suggesting it may contribute to observed H_2 production. We also conducted new proteomic and transcriptomic analysis to gain insights into the metabolic capabilities of 'Ca. L. ossiferum' (Table S4); this archaeon also synthesizes an [FeFe]-hydrogenase at high levels (18.03 LFQ intensity [log2]), but its [NiFe]-hydrogenase (24.00 LFQ intensity [log2]) is among the most abundant complexes in the cell and thus likely dominates H_2 production (Table S4; Figure S3). Nevertheless, the [FeFe]-hydrogenase may contribute to H_2 production given these enzymes typically have higher activities than their [NiFe] counterparts and all biochemically characterised group 3c [NiFe]-hydrogenases oxidise H_2 under cellular conditions (Greening et al. 2016; Costa et al. 2010; Kaster et al. 2011). Given the extremely slow growth rates and yields of both cultures (Imachi et al. 2020; Rodrigues-Oliveira et al. 2023) we are currently unable to conduct deeper analyses of the differential roles of these enzymes *in vivo*. Diverse [FeFe]-hydrogenases are also encoded by MAGs of several other Lokiarchaeia and Heimdallarchaeia (Figure 2.1; Table S1).

To better understand the structure and function of putative archaeal [FeFe]-hydrogenases, we performed structural modelling using AlphaFold2. Archaeal group A1 and E [FeFe]-hydrogenases are monomeric enzymes (HydA only) predicted to fold into a H-cluster domain with a solvent-exposed catalytic H-cluster (Figure 2.2A-B; Figure S4). However, in contrast to bacterial [FeFe]-hydrogenases, the loop structures of these enzymes are condensed and additional iron-sulfur clusters are absent (Figure S5 and S6). Each of the modelled enzymes contained the cysteine residues required to ligate the H-cluster, though differed in their proton-transferring residues (Figure S5). Thus, despite their small size, these enzymes are theoretically capable of H_2 catalysis. The group B [FeFe]-hydrogenase from 'Ca. P. syntrophicum' modelled as a heterodimer between the HydA and HydC subunits encoded by the gene cluster; it contains two 2×[4Fe-4S] ferredoxin-like domains, one that acts as an electron relay from the H-cluster, the other of unknown function separate from the main body of the protein (Figure 2.2C). Structural modelling also supported that archaeal group A3 [FeFe]-hydrogenases form trimeric electron-bifurcating complexes (HydABC) similar to those recently structurally characterised in fermentative and acetogenic bacteria (Figure 2.2D; Figure S7).

Archaeal [FeFe]-hydrogenases are catalytically active and display H-cluster spectroscopic signals

We tested whether group A1, B, and E [FeFe]-hydrogenases from archaea bind the catalytic H-cluster and produce H_2. To do so, we heterologously expressed five enzymes in E. coli BL21(DE3), anaerobically matured them using the synthetic mimic [^{2}Fe]adt ($[Fe_2(azadithiolate)(CO)_4(CN)_2]^{2-}$), and measured H_2 production of whole-cell lysates using gas chromatography as per established protocols (Table S3; Figure S8) (Berggren et al. 2013; Khanna et al. 2017; Land et al. 2019). H2 production was clearly discernible relative to negative controls in enzymes matured from all three groups (Figure 2.3A). The highest activity was observed from a Micrarchaeota group A1 enzyme (denoted Mu). Cell lysates containing Mu evolved H_2 at half the rate of the well-known [FeFe]-hydrogenase HydA1 from the green alga C. reinhardtii, which was included in all assays as a positive control (Land et al. 2019). The homologous enzyme from the groundwater archaeon 'Ca. Iainarchaeum andersonii' 76 (denoted Ia) also produced H_2, though at a hundredfold lower rate. We further observed substantial activity of the group B enzyme from the Asgard archaeon 'Ca. P. syntrophicum' (Imachi et al. 2020) (denoted Ps) and the group E enzyme from the groundwater archaeon 'Ca. Forterrea multitransposorum' (denoted Fm). We note that the observed H_2 production activities should not be considered as specific activities given potential variations in the efficiency of heterologous expression, folding, and maturation between the enzymes; this is illustrated by the differences in activity between the Mu and Ia hydrogenases despite their high degree of sequence homology. Nevertheless, these results validate that both DPANN and Asgard archaea encode functional [FeFe]-hydrogenases, and that the minimal group A1 and ultraminimal group E lineages are both active.

The group E [FeFe]-hydrogenase from 'Ca. F. multitransposorum' (Fm) was isolated as a representative example to provide more detailed insight into the properties of these ultra-minimalistic enzymes. Following purification under strictly anaerobic conditions and semi-enzymatic reconstitution of the iron-sulfur clusters, the iron content of the enzyme was determined to be 4.2 ± 0.4 per protein (Figure S9), in agreement with the structural modelling that this enzyme contains a single [4Fe4S] cluster (Figure 2.2). The reconstituted Fm hydrogenase was subsequently incubated with [2Fe]adt. Successful H-cluster assembly was verified through Attenuated Total Reflection Fourier transformed infrared (ATR-FTIR) spectroscopy, given sharp cofactor bands in the expected CO/CN ligand band region of the FTIR spectra were readily observed (Figure 2.3B; Figure S10). As elaborated in the Supplementary Note 3, our spectroscopic analysis suggested that the enzyme was isolated in an inhibited state. However, photochemical reduction resulted in the transition to catalytically active states (Figure 2.3B; Figure S10) reminiscent of the oxidised active ready states (Hox and HoxH 78) and a further reduced state.

The catalytic properties of Fm were studied by protein film electrochemistry (PFE). Cyclic voltammetry traces of the enzyme recorded under a H_2 atmosphere showed the typical bidirectional catalytic behaviour commonly associated with [FeFe]-hydrogenases 81 (Figure 2.3C). A comparison of the reducing and oxidising currents observed at high driving force (± 300 mV vs reversible hydrogen electrode, RHE) indicated that the enzyme is clearly biased towards H+ reduction catalysis relative to H_2 gas oxidation. This is in contrast to its closest characterized homolog, the group C [FeFe]-hydrogenase from the bacterium *Thermotoga maritima* (28% sequence identity), that showed a clear preference for H_2 oxidation (Chongdar et al. 2018). Collectively, these results show archaeal monomeric [FeFe]-hydrogenases binds a redox-active H-cluster in a very well-defined environment, and suggest they mediate fermentative H_2 production.

[FeFe]- and [NiFe]-hydrogenases associate into complexes in uncultivated archaea

Remarkably, the group F [FeFe]-hydrogenases appear to form complexes with [NiFe]-hydrogenases. 21 genomes encoded these complexes through five-gene clusters, including from the classes Bathyarchaeia, Brockarchaeia, Thermoplasmata, Thermoproteia, and Lokiarchaeia (Figure 2.1; Table S2). Their defining feature is the fusion of a C-terminal [FeFe]-hydrogenase catalytic domain (HydA) with an N-terminal domain homologous to the group 3 [NiFe]-hydrogenase small subunit (HyhS) (Figure 2.4A and Figure S4). Four other genes are contiguous with this fusion: the large subunit of the group 3 [NiFe]-hydrogenase (HyhL); the diaphorase (HydB) and thioredoxin (HydC) subunits of the electron-bifurcating/confurcating group A3 [FeFe]-hydrogenase; and a conduit subunit containing four iron-sulfur clusters (herein HydD) (Figure S2). The thioredoxin, diaphorase, and conduit subunits are respectively homologous to NuoE, NuoF, and NuoG that together form the NADH dehydrogenase module of complex I. All four genes are also present in the recently identified and structurally characterised electron-bifurcating [NiFe]-hydrogenase from Acetomicrobium mobile 14; however, the archaeal complex is distinct given it contains a true [FeFe]-hydrogenase catalytic subunit with a H-cluster domain, which is fused to the [NiFe]-hydrogenase small subunit. Of the archaeal [FeFe]-hydrogenases, the group F enzymes show the most conserved genetic organization in archaea, presumably due to their association with [NiFe]-hydrogenases (Figure S2). As elaborated in Supplementary Note 1, the genetically and phylogenetically distinct group G [FeFe]-hydrogenases exclusive to the Brockarchaeia genus JAAOZO01 are also likely to form hybrid complexes (Figure S2; Table S2). To confirm the catalytic activity of the group F [FeFe]-hydrogenase, we recombinantly expressed and artificially matured two HydA-HyhS fusion proteins from the candidate lineage Thermoplasmata SG8-5. One of these enzymes rapidly produced H_2 over the time course (Th1) with relative activities of 5.1% compared to the *C. reinhardtii* enzyme (Figure 2.4D; Table S5). In contrast, the other enzyme (Th2) exhibited only low levels of activity (Figure 2.4D). The proteins proved challenging to purify, preventing a detailed in vitro characterization. It is likely that these enzymes would display even higher activity if assembled into a complex with the other subunits, though this would be exceptionally challenging to achieve given the archaea encoding them are uncultivated and [NiFe]-hydrogenases are recalcitrant to heterologous production. These measurements nevertheless confirm these are bona fide hydrogenases and that their activities are attributable to the H-cluster.

Finally, we performed AlphaFold2 modelling to infer whether the [FeFe]- and [NiFe]-hydrogenases associate (Figure 4B-C). When modelled alone, the hybrid subunit was predicted to contain a [FeFe]-hydrogenase catalytic domain and a [NiFe]-hydrogenase iron-sulfur cluster domain separated by a long flexible linker (Figure S4). Conserved cysteine ligands required to ligate both the H-cluster of the [FeFe]-hydrogenase and the three electron-relaying [4Fe4S] clusters of the [NiFe]-hydrogenase component were both observed (Figure S5; Figure 2.4C; Supplementary Note 4). However, modelling of all five genetically contiguous subunits (HydA-HyhS, HydBCD, HyhL) suggests they form a stable electron-bifurcating/confurcating complex (ΔG = -97.06 to -219.7 kJ mol-1; PISA analysis) that associate through multiple hydrogen bonds and salt bridges (Figure S7). As elaborated in Supplementary Note 4, the structural model suggests the complex receives electrons through the [FeFe]-hydrogenase arm or the [NiFe]-hydrogenase via a series of iron-sulfur clusters to a probable electron-converging [4Fe4S] cluster on the hybrid subunit, or vice versa. Thereafter electrons are predicted to be simultaneously transferred to the high-potential NAD+ at the HydC subunit and an undetermined low-potential acceptor (likely ferredoxin) at the glutamate synthase (GltA) domain of the HydB subunit (Figure 2.4B). These observations are remarkable given [NiFe]- and [FeFe]-hydrogenases are not known to associate. Moreover, they suggest surprising modularity of the [NiFe]-hydrogenase small subunit given it has evidently co-evolved with both [NiFe]- and [FeFe]-hydrogenases.

[FeFe]-hydrogenases enable fermentation and electron-bifurcation in diverse archaea

We sought to understand the role of the various [FeFe]-hydrogenases in the metabolism of archaea. To do so, we annotated the high-quality archaeal genomes for genes associated with major energy conservation and carbon acquisition processes. All [FeFe]-hydrogenase-encoding archaea are predicted to be obligate anaerobes given they lacked terminal oxidases. This is consistent with the retrieval of the genomes from typically anoxic ecosystems, especially groundwater, anaerobic digesters, and sediments from hot springs, hydrothermal vents, and freshwater (Figure 2.1; Table S1). Based on the retrieved genomes, DPANN archaea are likely to be symbiotic obligate fermenters dependent on host-derived organic compounds; consistently, they often encoded genes for the degradation and fermentation of carbohydrates (primarily starch) and aromatic compounds, but generally lacked respiratory reductases or carbon fixation pathways (Figure 2.1). The other archaeal phyla were predicted to be capable of a wider range of metabolic strategies, in line with their larger genome sizes (Figure S1), including beta oxidation, anaerobic respiration, and carbon fixation to varying extents (Figure 2.1). Notably, some Asgard archaea encoded fumarate reductases (Frd), reductive dehalogenases (Rdh), and anaerobic sulfite reductases (Asr), with the latter enzyme also encoded in certain MAGs from four other phyla. Several lineages were also predicted to be capable of autotrophy through the Wood-Ljungdahl (Lokiarchaeia, Thermoproteota) or reverse tricarboxylic acid (Lokiarchaeia, Heimdallarchaeia, Thermoplasmatota) pathways (Figure 2.1; Table S5). Most of the MAGs also encoded RuBisCO lineages known to function in nucleoside salvage (Sato et al. 2007).

The group A, B, and E [FeFe]-hydrogenases likely facilitate cofactor regeneration during organic carbon fermentation in diverse archaea. Half of the archaea encode 2-oxoacid-ferredoxin oxidoreductases, such as pyruvate-ferredoxin oxidoreductases that couple the oxidation of the end product of glycolysis (pyruvate) to the reduction of ferredoxin, and most can gain ATP by converting the derived acetyl-CoA to the end product acetate via the acetyl-CoA synthetase or acetate kinase reaction (Figure 2.1). As supported by the structural modelling (Figure 2.2), the monomeric group A1, B, and E hydrogenases are predicted to couple ferredoxin reoxidation to H_2 production. Also consistent with this role, 2-oxoacid-ferredoxin oxidoreductase genes are frequently adjacent to [FeFe]-hydrogenase genes in Nanoarchaeota and Micrarchaeota MAGs (Figure S2). By contrast, the trimeric group A3 [FeFe]-hydrogenases are predicted to simultaneously reoxidize ferredoxin (reduced primarily by pyruvate-ferredoxin oxidoreductase) and NADH (e.g. reduced during glycolysis) (Figure 2.2), in line with their bacterial counterparts (Schut and Adams 2009; Furlan et al. 2022; Katsyv et al. 2023). Congruently, the electron-bifurcating/confurcating group A3 [FeFe]-hydrogenases are associated with those DPANN archaea harbouring relatively complex carbohydrate degradation pathways (i.e. Woesearchaeles, Altarchaeota, some Iainarchaeota) through which both NAD+ and ferredoxin will be reduced (Figure 2.1). It is also notable that many DPANN archaea encode the minimal group A1 and ultraminimal group E [FeFe]-hydrogenases as their sole H2-metabolising enzymes (Table S6); in conjunction with the simpler maturation pathways of [FeFe]-hydrogenases compared to [NiFe]-hydrogenases, these genome-reduced microorganisms (average completeness-normalized genome size: 1.2 Mbp) have minimised the genetic and cellular costs of metabolising H_2. Further studies are needed to determine what biochemical features differentiate the monomeric group A1, B, and

E [FeFe]-hydrogenases and whether they have distinct physiological roles. It cannot be ruled out that the group B enzymes instead consume H_2 to support anaerobic respiration or carbon fixation in Asgard archaea; however, this seems unlikely given their reported H_2-evolving activities (Figure 2.3), structural features (Figure 2.2), and the hydrogenogenic lifestyle of 'Ca. P. syntrophicum' (Imachi et al. 2020).

The physiological role of the hybrid hydrogenases is unclear. These enzymes are exclusively encoded by more metabolically versatile archaea, including those with the capacity for carbon fixation, beta oxidation, and energy conversion using group 4 [NiFe]-hydrogenases (Figure 2.1). The predicted structures suggest that that these enzymes contribute to electron bifurcation by transferring electrons from H_2 to both NAD and likely ferredoxin, or vice versa (Figure 3). However, it is peculiar that a single complex contains two seemingly redundant hydrogenase modules. A potential explanation is that one of the hydrogenase modules might act on a substrate other than H_2, especially given group 3 [NiFe]-hydrogenases can serve as sulfhydrogenases in vitro (i.e. mediating reduction of elemental sulfur to hydrogen sulfide) (Ma et al. 1993). A further rationale is that the two hydrogenase modules may differ in their affinities and/or oxygen tolerance, enabling efficient H_2 oxidation across a wide range of environmental conditions. However, perhaps the strongest possibility is that the complexes act as a redox valves, transferring electrons either from NADH and reduced ferredoxin to a H_2-evolving [FeFe]-hydrogenase when reductant accumulates, and from a H_2-consuming [NiFe]-hydrogenase to NAD and oxidised ferredoxin otherwise; this is consistent with previous observations that electron-bifurcating hydrogenases act as H_2-producing redox valves during heterotrophic growth of bacteria in response to variations in substrate availability and redox state (Wang et al. 2013). Other enzyme complexes regulating opposite reactions have recently been reported, namely between glutamate synthase (GltAB) and glutamate dehydrogenase (GudB) (Jayaraman et al. 2022), with the GltA domain of these enzymes also shared in the hybrid hydrogenase complex.

[FeFe]-hydrogenases have been acquired by archaea on multiple occasions and have an ancient association with [NiFe]-hydrogenases

We investigated the evolutionary history of [FeFe]-hydrogenases through phylogenetic analysis of its catalytic subunit (Figure 2.5) and three maturases. In agreement with their classification into known groups, the archaeal group A1, A3, and B [FeFe]-hydrogenases clustered with various bacterial and eukaryotic homologs; the group A enzymes form at least six radiations, suggesting they were laterally acquired from bacterial enzymes over several independent events, whereas the archaeal group B and E enzymes each formed a monophyletic clade (Figure 2.5). Based on phylogenies constructed using the best supported model (Figure 2.5) and their structural simplicity (Figure 2.2B), the ultraminimal fermentative group E enzymes of archaea form a sister clade to the multidomain sensory group C hydrogenases of bacteria (Greening et al. 2016; Chongdar et al. 2018; Greening et al. 2019; Zheng et al. 2014). We also re-examined the putative origin of [FeFe]-hydrogenases in eukaryotes, in light of the syntrophy hypotheses for eukaryogenesis (Martin and Müller 1998; Moreira and Lopez-Garcia 1998; Sousa et al. 2016; Spang et al. 2019; Imachi et al. 2020), given our finding [FeFe]-hydrogenases are present both in Asgard archaea and unicellular eukaryotes. Our data do not support the hypothesis that all eukaryotic [FeFe]-hydrogenases were vertically acquired from an Asgard archaeal ancestor. The eukaryotic enzymes cluster into at least five disparate clades, spanning both group A and B, suggesting multiple horizontal acquisitions (Figure 2.5). In addition, [FeFe]-hydrogenases are absent from the Asgard genomes most related to eukaryotes (Figure 2.1) and the alphaproteobacterial genomes most related to mitochondria

(Greening et al. 2016; Stairs et al. 2021; Stairs et al. 2020), though this view could change with the addition of new genomes. Nevertheless, archaeal and eukaryotic [FeFe]-hydrogenases clustered together in three of these lineages (including Lokiarchaeia with Tritrichomonas), suggesting some eukaryotes and archaea may have laterally exchanged [FeFe]-hydrogenase genes during their diversification (Figure 2.5).

We also studied the distribution and phylogeny of the three maturases that synthesise the [FeFe]-hydrogenase cofactor, HydE, HydF, and HydG. Of the 130 archaeal genomes encoding [FeFe]-hydrogenases, just five encode a full set of maturases and three others encode an incomplete set (Table S3). Though genome incompleteness means that the co-occurrence of hydrogenases and maturases will be underestimated, this does not explain the absence of maturases in most genomes. Indeed, maturase genes are even absent in the genome of the cultured hydrogenogenic archaeon 'Ca. P. syntrophicum' and the closed genome of 'Ca. F. multitransposorum' from which an active [FeFe]-hydrogenase was heterologously produced (Figure 2.3). In turn, it is likely that archaea synthesise [FeFe]-hydrogenases through an alternative pathway. The existence of an alternative pathway is consistent with reports that various eukaryotes lacking all (Giardia, Entamoeba (Nixon et al. 2003; Lloyd et al. 2002)) or some (Mastigamoeba (Nývltová et al. 2013)) maturases still make catalytically active H_2-producing hydrogenases, as well as recent reports of a cytosolic [FeFe]-hydrogenase in *Trichomonas vaginalis* (Smutná et al. 2022) and the heterologous production of active [FeFe]-hydrogenases in Synechocystis cells lacking maturases (Berto et al. 2011). In line with findings for the structural subunits, phylogenetic analysis of HydE, HydF, and HydG suggests that archaea acquired maturases on several occasions. Several archaeal maturases are closely related to those that we recently identified in Chlamydiae and eukaryotes(Smutná et al. 2022).

Phylogenetic analyses also suggest an ancient origin of the hybrid hydrogenases. In phylogenetic trees of the [FeFe]-hydrogenase catalytic subunit, sequences of group F and G hydrogenases together formed a monophyletic branch distant from the group A and group B to E superclades, suggesting an early divergence of these archaeal enzymes (Figure 2.5). These sequences formed long branches, given their divergence from other hydrogenases (~25% sequence identity to the model hydrogenase *C. reinhardtii* HydA1). The discovery of these divergent archaeal hydrogenases raises the question of whether [FeFe]-hydrogenase first evolved in archaea or bacteria; however, the absence of non-hydrogenase homologs ancestral to [FeFe]-hydrogenases precludes confident rooting of the tree needed to make strong inferences. In phylogenetic trees of [NiFe]-hydrogenases, the catalytic (large) subunits of the fusion proteins formed multiple clusters In contrast, the iron-sulfur (small) domain fused to the group F [FeFe]-hydrogenase and the iron-sulfur subunit downstream of the group G [FeFe]-hydrogenase each form distinct monophyletic subgroups within the group 3 [NiFe]-hydrogenases (Figure S16). Thus, as dictated by the gene fusion, the iron-sulfur domain of hybrid complexes more strongly co-evolved with the [FeFe]- than [NiFe]-hydrogenase catalytic subunits. We propose that the components of the hybrid hydrogenases are formally recognised as distinct lineages, namely the group F and G [FeFe]-hydrogenases and group 3f and 3g [NiFe]-hydrogenases, given their distinct phylogenies, structures, and potential physiological roles.

2.4 Conclusion

Archaea have evolved remarkably disparate ways to use [FeFe]-hydrogenases to adapt to anaerobic environments. On one hand, DPANN archaea have evolved ultraminimal enzymes to efficiently dispose of reductant derived from carbohydrate fermentation. Conversely, lineages such as Brockarchaeia and Lokiarchaeia use unique hybrids of [NiFe]- and [FeFe]-hydrogenases – among the most complex hydrogenases described to date – to support their diverse redox biology. These discoveries expand the [FeFe]-hydrogenases from four to seven groups each with distinct phylogenies, structures, and functions. In addition to increasing understanding of archaeal biology, these findings also redefine our understanding of hydrogen metabolism and hydrogenase biochemistry by revealing: (i) [FeFe]-hydrogenases from all three domains of life are active, (ii) the two dominant hydrogenase classes have co-evolved, and (iii) a new size minimum for hydrogenases. The ultraminimal hydrogenases also provide ideal templates both to understand enzymatic H_2 catalysis, for example determinants of rate, directionality, affinity, and oxygen sensitivity, as well as flexible scaffolds for directed evolution of efficient H_2-converting biocatalysts. More broadly, our approach also emphasizes the potential of combining genome-resolved metagenomics with accurate protein structure prediction and heterologous production studies to discover new enzymes and functions in uncultured microorganisms.

2.5 Figures

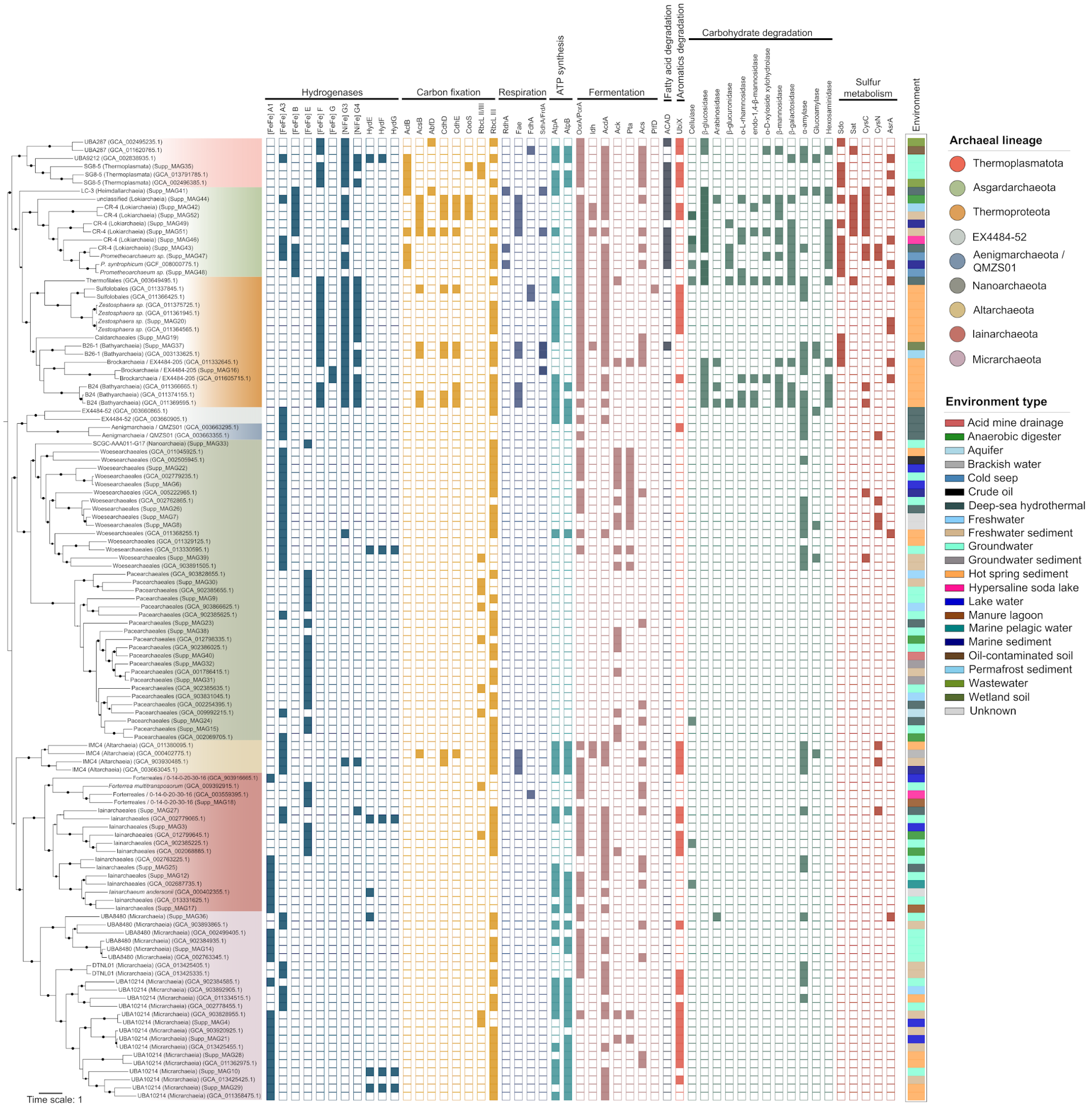


Figure 2.1 Phylogenetically and metabolically diverse archaea encode [FeFe]-hydrogenases. The left portion of the figure shows a maximum-likelihood phylogenomic tree (model LG+F+G4) based on the concatenated 15 ribosomal marker proteins of archaeal genomes that encode [FeFe]-hydrogenases. Results are shown for the 118 (out of 130) genomes that are at least 60% complete, less than 5% contaminated, and contain at least 75% of the 15 syntenic proteins. Branches are colour-coded encoding according to the respective phylum. Black circles indicate bootstrap support values over 80%. The middle portion shows the presence of key metabolic genes involved in different metabolic processes. Carbon fixation: ATPcitrate lyase beta-subunit (AclB), acetyl-CoA synthase beta subunit (AcsB), propionyl-CoA synthetase (PrpE), 4-hydroxybutyryl-CoA dehydratase / vinylacetyl-CoA-deltaisomerase (AbfD), CODH/ACS complex subunit delta (CdhD), CODH/ACS complex subunit gamma (CdhE), anaerobic carbon monoxide dehydrogenase catalytic subunit (CooS), type II/III ribulose-bisphosphate carboxylase (RbcL II/II), type III

ribulosebisphosphate carboxylase (RbcL III); respiration: reductive dehalogenase (RdhA), formaldehyde activating enzyme (Fae), glutathione-independent formaldehyde dehydrogenase (FdhA), the reversible succinate dehydrogenase and fumarate reductase flavoprotein (SdhA/FrdA); ATP synthesis: ATP synthase subunit alpha (AtpA), ATP synthase subunit beta (AtpB); fermentation: 2-oxoacid:ferredoxin or pyruvate:ferredoxin oxidoreductase alpha subunit (PorA/OorA), isocitrate dehydrogenase (Idh), ADP-forming acetyl-CoA synthetase (AcdA), acetate kinase (Ack), phosphate acetyltransferase (Pta), acetyl-CoA synthetase (Acs), formate Cacetyltransferase (PflD); fatty acid degradation: acyl-CoA dehydrogenase (ACAD); aromatics degradation: flavin prenyltransferase (UbiX); sulfur metabolism: sulfur dioxygenase (Sdo), sulfate adenylyltransferase (Sat), adenylylsulfate kinase (CysC), sulfate adenylyltransferase subunit 1 (CysN), anaerobic sulfite reductase subunit A (AsrA). The right portion shows the diverse environments the archaeal genomes were retrieved from. Note that the phylum QMZS01 was classified as Aenigmatarchaeota in GTDB R06-RS207 while Thermoproteota class EX4484−205 was proposed as Brockarchaeia.

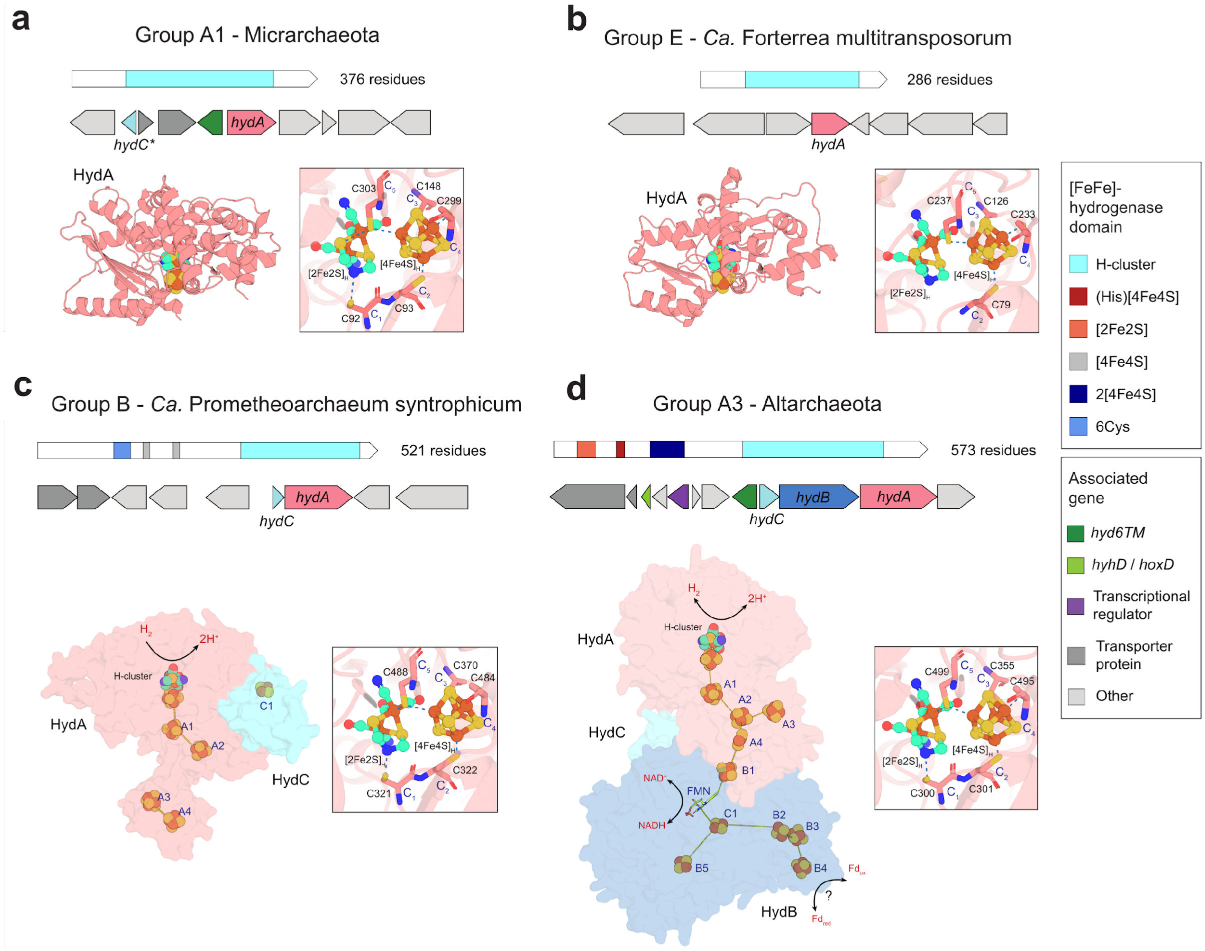


Figure 2.2 Archaea encode genetically and structurally diverse [FeFe]-hydrogenases. Catalytic domain structure, genetic organisation, and AlphaFold2-based structural modelling of representative [FeFe]-hydrogenases encoded in archaeal genomes. A) Group A1 [FeFe]-hydrogenase from UBA95 sp002499405 (Micrarchaeota). B) Group E [FeFe]-hydrogenase from '*Ca.* Forterrea multitransposorum'. C) Group B [FeFe]-hydrogenase complex from '*Ca.* Prometheoarchaeum syntrophicum'. D) Group A3 [FeFe]-hydrogenase complex from DSAL01 sp011380095 (Altarchaeota). For each panel, the catalytic domain (H-cluster), iron-sulfur binding motifs, and amino acid sequence length are shown at the top. Genes encoding hydrogenase structural subunits are shown in their genetic context are shown beneath, labelled and coloured consistent with the corresponding subunit in the structural models. Predicted cofactors are positioned based on the structures of homologous proteins. A zoomed view of the H-cluster and conserved coordinating cysteine residues (C_1 to C_5) is shown for each group. For group B and A3 enzymes, FeS clusters within plausible electron transfer distance are connected by dashed lines. *hydA*, [FeFe]-hydrogenase; *hydB*, diaphorase; *hydC*, thioredoxin; *hydD*, nuoG-like conduit protein; *hyd6TM*, uncharacterised 4 to 6-helix transmembrane protein associated with group A [FeFe]-hydrogenases; (His)[4Fe4S], $(Cys)_3$His-ligated [4Fe4S] cluster binding domain; [2Fe2S], [2Fe2S] cluster binding domain; [4Fe4S], [4Fe4S] cluster binding domain; 2[4Fe4S], bacterial ferredoxin-like 2[4Fe4S] cluster binding domain; 6Cys, putative iron-sulfur cluster binding domain. * HydC protein in group A1 gene cluster was not predicted to form a complex with HydA. Surface structures are used for the multisubunit group B and A3 [FeFe]-hydrogenases, with ribbon diagram versions provided in Figure S4.

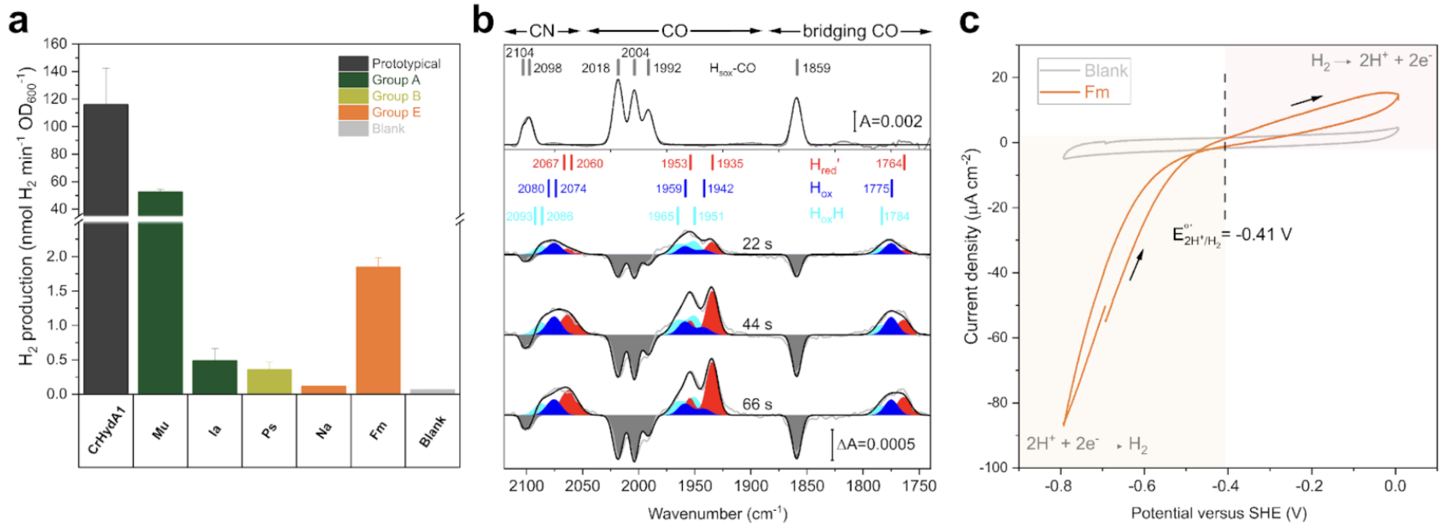

a
H$_2$ production (nmol H$_2$ min^{-1} OD$_{600}$$^{-1}$)
160
140
120
100
80
60
40
2.0
1.5
1.0
0.5
0.0
Prototypical
Group A
Group B
Group E
Blank
CrHydA1
Mu
Ia
Ps
Na
Fm
Blank
b
CN
CO
bridging CO
2104
2098
2018
2004
1992
H$_{sox}$-CO
1859
A=0.002
2067
2060
1953
1935
H$_{red}$'
1764
2080
2074
1959
1942
H$_{ox}$
1775
2093
2086
1965
1951
H$_{ox}$H
1784
22 s
44 s
66 s
ΔA=0.0005
2100
2050
2000
1950
1900
1850
1800
1750
Wavenumber (cm^{-1})
c
Blank
Fm
H$_2$ → 2H$^+$ + 2e$^-$
$E^{o'}_{2H^+/H_2}$ = -0.41 V
2H$^+$ + 2e$^-$ → H$_2$
Current density (μA cm^{-2})
20
0
-20
-40
-60
-80
-100
-0.8
-0.6
-0.4
-0.2
0.0
Potential versus SHE (V)

Figure 2.3 Three classes of [FeFe]-hydrogenases encoded by archaea are catalytically active. A) H_2 gas production monitored from cell lysates in E. coli BL21(DE3) cells expressing group A1, B, and E [FeFe]-hydrogenases from archaea. All cell lysates including the blank were activated by addition of [2Fe]adt. H2 was measured by GC after addition of methyl viologen and dithionite to activated cell lysates, set to pH 6.8 with 100 mM KPi buffer. Activities are normalized for number of cells used (nmol H_2 min-1 OD_{600}-1) and error bars reflect standard deviation from biological triplicates. The strain expressing prototypical CrHydA1 was used as a positive control while "Blank" represents the same strain but containing an empty vector. B) FTIR spectra of the group E [FeFe]-hydrogenase from 'Ca. Forterrea multitransposorum' (Fm) after heterologous expression, semisynthetic maturation with [2Fe]adt, and purification. The absorbance spectrum (top) indicates a CO inhibited di-ferrous H-cluster state (Hsox-CO). The difference spectrum (bottom) illustrates the transitions of Fm into catalytically active states through photoreduction (illumination after the addition of eosin Y as a photosensitizer and triethanolamine as a sacrificial electron donor). During illumination bands associated with the highly oxidized CO-inhibited state decreased (grey bands), while new bands reflecting reduced and catalytically active H-cluster states appear, assigned to HoxH (cyan), Hox (blue) and Hred (red) (spectra arranged chronologically from top to bottom). C) Cyclic voltammetry traces of immobilized Fm (orange) with H2 oxidation current densities at high potentials and H+ reduction currents at low potentials. The 2H+/H2 redox couple potential is indicated with a dashed line (E0′2H+/H_2). Scan direction is indicated by black arrows. The "Blank" trace (grey) represents the electrode without an immobilized enzyme film. The experiments were performed on two independent films for each enzyme at pH 7.0 (5 mM MES, 5 mM CHES, 5 mM HEPES, 5 mM TAPS, 5 mM NaOAc, 0.1 M Na_2SO_4) and under 1 atm H_2.

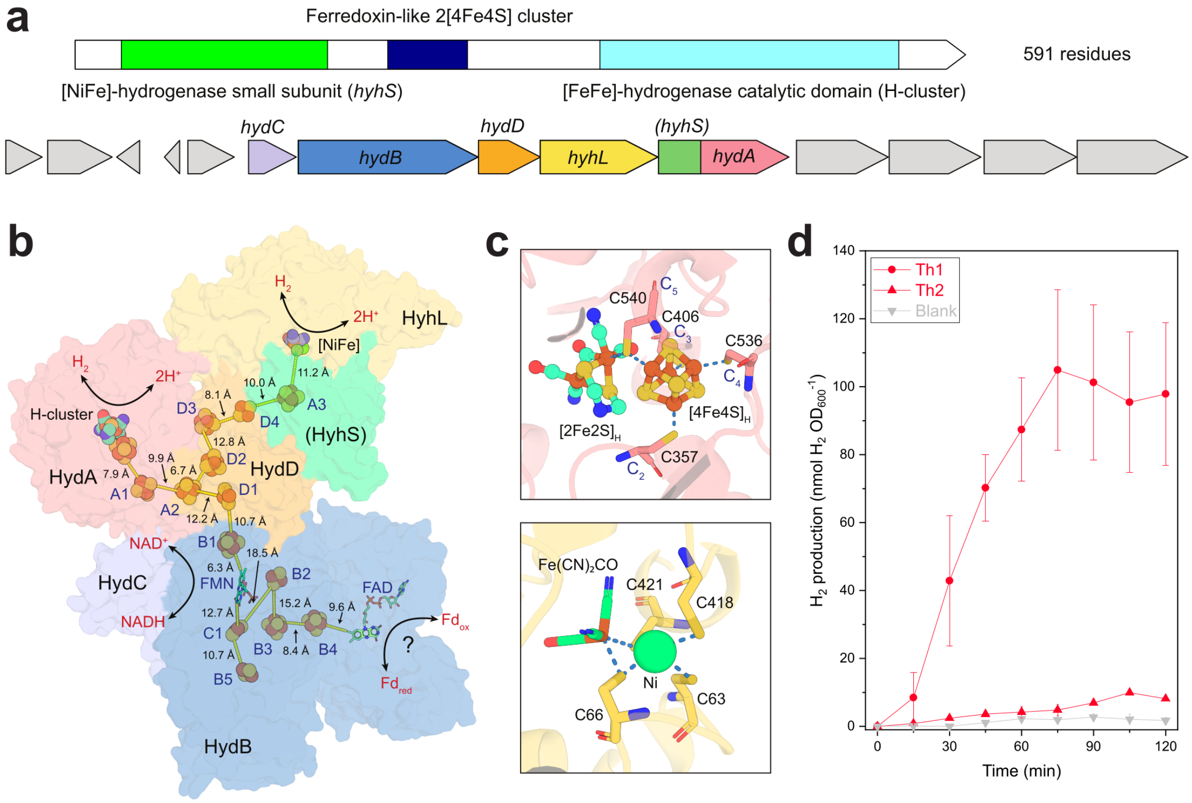


Figure 2.4 [FeFe]- and [NiFe]-hydrogenases encoded by archaea are predicted to form unique complexes. A) Catalytic domain structure and predicted operon encoding a putative complex of a group F [FeFe]-hydrogenase and group 3 [NiFe]-hydrogenase in Thermoplasmatota UBA147 sp002496385. *hydA*, [FeFe]-hydrogenase; *hydB*, diaphorase; *hydC*, thioredoxin; *hydD*, *nuoG*-like conduit protein; *hyhL*, group 3 [NiFe]-hydrogenase catalytic subunit; *hyhS*, group 3 [NiFe]-hydrogenase small subunit. B) Predicted surface structure, cofactor composition, and electron flow through four potential arms in the hybrid hydrogenase complex. [FeS] clusters are numbered and labelled according to their subunit of origin (e.g. A1, A2, A3 originate from the HydA subunit). C) Atomic structure of the predicted [FeFe]- and [NiFe]-hydrogenase active sites in the hybrid enzyme. Distances between catalytic cluster and coordinating residues of less than 2.5 Å are shown as blue dotted lines. D) H_2 gas production monitored from cell lysates in *E. coli* BL21(DE3) cells expressing the Th1 and Th2 [FeFe]-hydrogenases from archaea. The cell lysates were activated by addition of $[2Fe]^{adt}$. H_2 levels were measured every 15 mins for 2 hours by gas chromatography after addition of methyl viologen and dithionite to activated cell lysates, set to pH 6.8 with 100 mM KPi buffer. Activities are normalized for number of cells used (nmol H_2 OD_{600}^{-1}) and error bars reflect standard deviations from two biological triplicates. "Blank" represents the same strain but containing an empty vector.

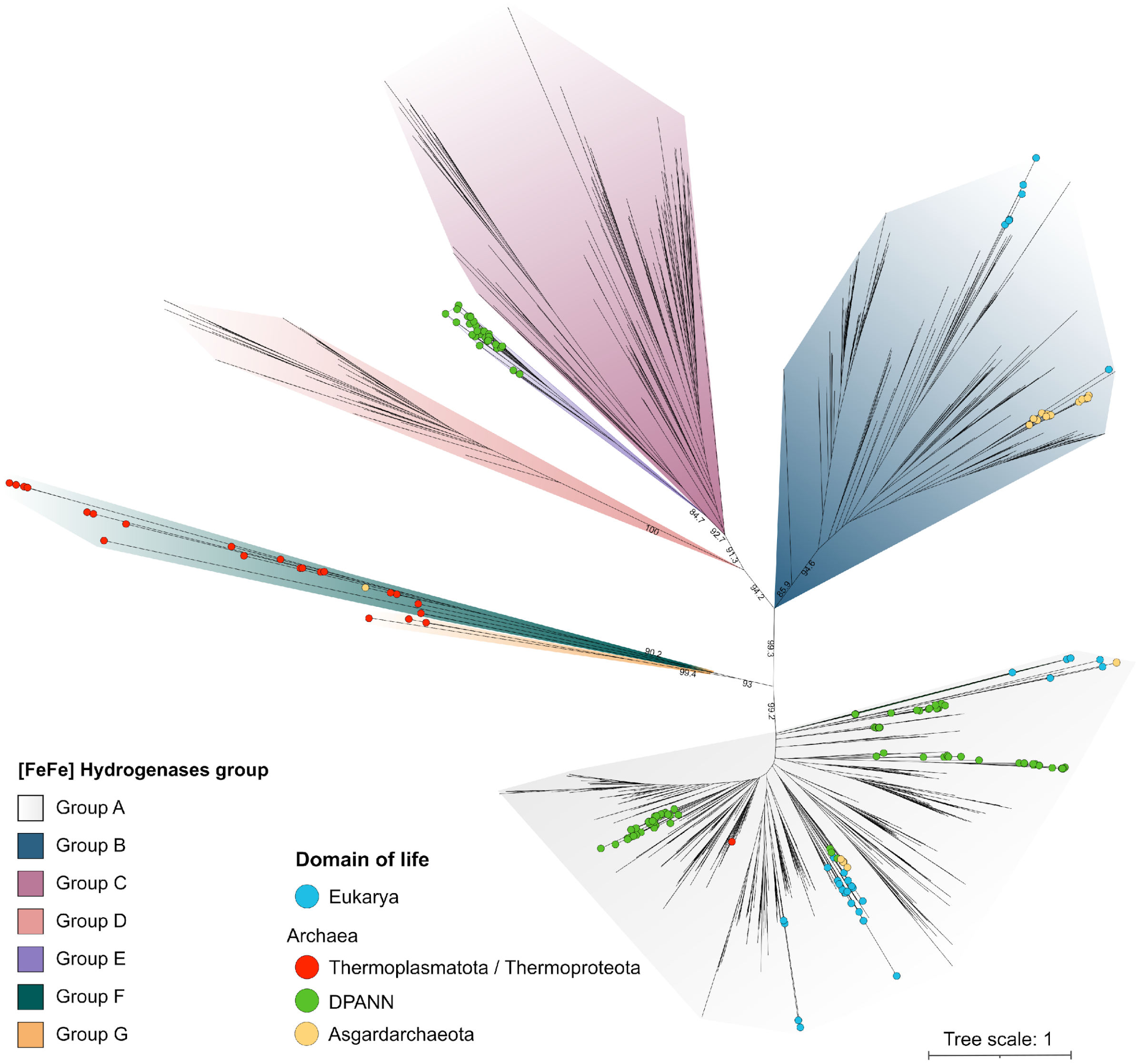


Figure 2.5 [FeFe]-hydrogenases are diverse and ancient in archaea. An unrooted maximum-likelihood phylogenetic tree of the catalytic subunit (HydA) of [FeFe]-hydrogenases and novel hybrid hydrogenases. The tree was constructed based on 3,677 amino acid sequences using the LG+C20+R+F model. The numbers at the branches indicate the aLRT (Approximate Likelihood Ratio Test) support values. Coloured circles at the tip indicate sequences from eukaryotes and major archaeal groups.

2.6 References

Al-Shayeb, B., Schoelmerich, M. C., West-Roberts, J., Valentin-Alvarado, L. E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M. J., Williams, K. H., Doudna, J. A., & Banfield, J. F. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature*, *610*(7933), 731–736.

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., Thomas, B. C., Singh, A., Wilkins, M. J., Karaoz, U., Brodie, E. L., Williams, K. H., Hubbard, S. S., & Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 13219.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* , *36*(7), 2251–2252.

Benoit, S. L., Maier, R. J., Sawers, R. G., & Greening, C. (2020). Molecular Hydrogen Metabolism: a Widespread Trait of Pathogenic Bacteria and Protists. *Microbiology and Molecular Biology Reviews: MMBR*, *84*(1). https://doi.org/10.1128/MMBR.00092-19

Berggren, G., Adamska, A., Lambertz, C., Simmons, T. R., Esselborn, J., Atta, M., Gambarelli, S., Mouesca, J. M., Reijerse, E., Lubitz, W., Happe, T., Artero, V., & Fontecave, M. (2013). Biomimetic assembly and activation of [FeFe]-hydrogenases. *Nature*, *499*(7456), 66–69.

Berto, P., D'Adamo, S., Bergantino, E., Vallese, F., Giacometti, G. M., & Costantini, P. (2011). The cyanobacterium Synechocystis sp. PCC 6803 is able to express an active [FeFe]-hydrogenase without additional maturation proteins. *Biochemical and Biophysical Research Communications*, *405*(4), 678–683.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* , *30*(15), 2114–2120.

Borrel, G., Adam, P. S., McKay, L. J., Chen, L.-X., Sierra-García, I. N., Sieber, C. M. K., Letourneur, Q., Ghozlane, A., Andersen, G. L., Li, W.-J., Hallam, S. J., Muyzer, G., de Oliveira, V. M., Inskeep, W. P., Banfield, J. F., & Gribaldo, S. (2019). Wide diversity of methane and short-chain alkane metabolisms in uncultured archaea. *Nature Microbiology*, *4*(4), 603–613.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.

Buckel, W., & Thauer, R. K. (2013). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1827*(2), 94–113.

Calusinska, M., Happe, T., Joris, B., & Wilmotte, A. (2010). The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. *Microbiology*, *156*(Pt 6), 1575–1588.

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

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* , *25*(15), 1972–1973.

Carbonero, F., Benefiel, A. C., & Gaskins, H. R. (2012). Contributions of the microbial hydrogen economy to colonic homeostasis. *Nature Reviews. Gastroenterology & Hepatology*, *9*(9), 504–518.

Castelle, C. J., & Banfield, J. F. (2018). Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell*, *172*(6), 1181–1197.

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., & Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews. Microbiology*, *16*(10), 629–645.

Castelle, C. J., Méheust, R., Jaffe, A. L., Seitz, K., Gong, X., Baker, B. J., & Banfield, J. F. (2021). Protein Family Content Uncovers Lineage Relationships and Bacterial Pathway Maintenance Mechanisms in DPANN Archaea. *Frontiers in Microbiology*, *12*, 660052.

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., Frischkorn, K. R., Tringe, S. G., Singh, A., Markillie, L. M., Taylor, R. C., Williams, K. H., & Banfield, J. F. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology: CB*, *25*(6), 690–701.

Chen, S.-C., Musat, N., Lechtenfeld, O. J., Paschke, H., Schmidt, M., Said, N., Popp, D., Calabrese, F., Stryhanyuk, H., Jaekel, U., Zhu, Y.-G., Joye, S. B., Richnow, H.-H., Widdel, F., & Musat, F. (2019). Anaerobic oxidation of ethane by archaea from a marine hydrocarbon seep. *Nature*, *568*(7750), 108–111.

Chongdar, N., Birrell, J. A., Pawlak, K., Sommer, C., Reijerse, E. J., Rüdiger, O., Lubitz, W., & Ogata, H. (2018). Unique Spectroscopic Properties of the H-Cluster in a Putative Sensory [FeFe] Hydrogenase. *Journal of the American Chemical Society*, *140*(3), 1057–1068.

Constant, P., Poissant, L., & Villemur, R. (2009). Tropospheric H2 budget and the response of its soil uptake under the changing environment. *The Science of the Total Environment*, *407*(6), 1809–1823.

Costa, K. C., Wong, P. M., Wang, T., Lie, T. J., Dodsworth, J. A., Swanson, I., Burn, J. A., Hackett, M., & Leigh, J. A. (2010). Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(24), 11050–11055.

Cracknell, J. A., Vincent, K. A., & Armstrong, F. A. (2008). Enzymes as working or inspirational electrocatalysts for fuel cells and electrolysis. *Chemical Reviews*, *108*(7), 2439–2461.

De Anda, V., Chen, L.-X., Dombrowski, N., Hua, Z.-S., Jiang, H.-C., Banfield, J. F., Li, W.-J., & Baker, B. J. (2021). Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. *Nature Communications*, *12*(1), 2404.

Dietrich, H. M., Righetto, R. D., Kumar, A., Wietrzynski, W., Trischler, R., Schuller, S. K., Wagner, J., Schwarz, F. M., Engel, B. D., Müller, V., & Schuller, J. M. (2022). Membrane-anchored HDCR nanowires drive hydrogen-powered CO2 fixation. *Nature*, *607*(7920), 823–830.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319.

Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P., & Spang, A. (2019). Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiology Letters*, *366*(2). https://doi.org/10.1093/femsle/fnz008

Dong, X., Greening, C., Rattray, J. E., Chakraborty, A., Chuvochina, M., Mayumi, D., Dolfing, J., Li, C., Brooks, J. M., Bernard, B. B., Groves, R. A., Lewis, I. A., & Hubert, C. R. J. (2019). Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature Communications*, *10*(1), 1816.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195.

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), 992–999.

Evans, P. N., Boyd, J. A., Leu, A. O., Woodcroft, B. J., Parks, D. H., Hugenholtz, P., & Tyson, G. W. (2019). An evolving view of methane metabolism in the Archaea. *Nature Reviews. Microbiology*, *17*(4), 219–232.

Evans, R. M., Siritanaratkul, B., Megarity, C. F., Pandey, K., Esterle, T. F., Badiani, S., & Armstrong, F. A. (2019). The value of enzymes in solar fuels research - efficient electrocatalysts through evolution. *Chemical Society Reviews*, *48*(7), 2039–2052.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2022). Protein complex prediction with AlphaFold-Multimer. In *bioRxiv* (p. 2021.10.04.463034). https://doi.org/10.1101/2021.10.04.463034

Feng, X., Schut, G. J., Haja, D. K., Adams, M. W. W., & Li, H. (2022). Structure and electron transfer pathways of an electron-bifurcating NiFe-hydrogenase. *Science Advances*, *8*(8), eabm7546.

Furlan, C., Chongdar, N., Gupta, P., Lubitz, W., Ogata, H., Blaza, J. N., & Birrell, J. A. (2022). Structural insight on the mechanism of an electron-bifurcating [FeFe] hydrogenase. *eLife*, *11*. https://doi.org/10.7554/eLife.79361

Greening, C., Biswas, A., Carere, C. R., Jackson, C. J., Taylor, M. C., Stott, M. B., Cook, G. M., & Morales, S. E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *The ISME Journal*, *10*(3), 761–777.

Greening, C., Geier, R., Wang, C., Woods, L. C., Morales, S. E., McDonald, M. J., Rushton-Green, R., Morgan, X. C., Koike, S., Leahy, S. C., Kelly, W. J., Cann, I., Attwood, G. T., Cook, G. M., & Mackie, R. I. (2019). Diverse hydrogen production and consumption pathways influence methane production in ruminants. *The ISME Journal*, *13*(10), 2617–2632.

Greening Chris, Ahmed F. Hafna, Mohamed A. Elaaf, Lee Brendon M., Pandey Gunjan, Warden Andrew C., Scott Colin, Oakeshott John G., Taylor Matthew C., & Jackson Colin J. (2016). Physiology, Biochemistry, and Applications of F420- and Fo-Dependent Redox Reactions. *Microbiology and Molecular Biology Reviews: MMBR*, *80*(2), 451–493.

Greening, C., Islam, Z. F., & Bay, S. K. (2022). Hydrogen is a major lifeline for aerobic bacteria. *Trends in Microbiology*, *30*(4), 330–337.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321.

He, C., Keren, R., Whittaker, M. L., Farag, I. F., Doudna, J. A., Cate, J. H. D., & Banfield, J. F. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nature Microbiology*, *6*(3), 354–365.

Holm, L., & Laakso, L. M. (2016). Dali server update. *Nucleic Acids Research*, *44*(W1), W351–W355.

Huang, W.-C., Liu, Y., Zhang, X., Zhang, C.-J., Zou, D., Zheng, S., Xu, W., Luo, Z., Liu, F., & Li, M. (2021). Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota. *Nature Communications*, *12*(1), 5281.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., … Takai, K. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*, *577*(7791), 519–525.

Jaffe, A. L., Bardot, C., Le Jeune, A.-H., Liu, J., Colombet, J., Perrière, F., Billard, H., Castelle, C. J., Lehours, A.-C., & Banfield, J. F. (2023). Variable impact of geochemical gradients on the functional potential of bacteria, archaea, and phages from the permanently stratified Lac Pavin. *Microbiome*, *11*(1), 14.

Jayaraman, V., Lee, D. J., Elad, N., Vimer, S., Sharon, M., Fraser, J. S., & Tawfik, D. S. (2022). A counter-enzyme complex regulates glutamate metabolism in Bacillus subtilis. *Nature Chemical Biology*, *18*(2), 161–170.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kang, D. D., 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*, *7*, e7359.

Kaster, A.-K., Moll, J., Parey, K., & Thauer, R. K. (2011). Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2981–2986.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Katsyv, A., Kumar, A., Saura, P., Pöverlein, M. C., Freibert, S. A., T Stripp, S., Jain, S., Gamiz-Hernandez, A. P., Kaila, V. R. I., Müller, V., & Schuller, J. M. (2023). Molecular Basis of the Electron Bifurcation Mechanism in the [FeFe]-Hydrogenase Complex HydABC. *Journal of the American Chemical Society*, *145*(10), 5696–5709.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , *28*(12), 1647–1649.

Khanna, N., Esmieu, C., Mészáros, L. S., Lindblad, P., & Berggren, G. (2017). In vivo activation of an [FeFe] hydrogenase using synthetic cofactors. *Energy & Environmental Science*, *10*(7), 1563–1567.

Kleinhaus, J. T., Wittkamp, F., Yadav, S., Siegmund, D., & Apfel, U.-P. (2021). [FeFe]-Hydrogenases: maturation and reactivity of enzymatic systems and overview of biomimetic models. *Chemical Society Reviews*, *50*(3), 1668–1784.

Krissinel, E. (2015). Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Research*, *43*(W1), W314–W319.

Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, *372*(3), 774–797.

Land, H., Ceccaldi, P., Mészáros, L. S., Lorenzi, M., Redman, H. J., Senger, M., Stripp, S. T., & Berggren, G. (2019). Discovery of novel [FeFe]-hydrogenases for biocatalytic H2-production. *Chemical Science* , *10*(43), 9941–9948.

Land, H., Senger, M., Berggren, G., & Stripp, S. T. (2020). Current State of [FeFe]-Hydrogenase Research: Biodiversity and Spectroscopic Investigations. *ACS Catalysis*, *10*(13), 7069–7086.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Laso-Pérez, R., Wegener, G., Knittel, K., Widdel, F., Harding, K. J., Krukenberg, V., Meier, D. V., Richter, M., Tegetmeyer, H. E., Riedel, D., Richnow, H.-H., Adrian, L., Reemtsma, T., Lechtenfeld, O. J., & Musat, F. (2016). Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature*, *539*(7629), 396–401.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296.

Leung, P. M., Grinter, R., Tudor-Matthew, E., Jimenez, L., Lee, H., Milton, M., Hanchapola, I., Tanuwidjaya, E., Peach, H. A., Carere, C. R., Stott, M. B., Schittenhelm, R. B., & Greening, C. (2022). Atmospheric hydrogen oxidation extends to the domain archaea. In *bioRxiv* (p. 2022.12.13.520232). https://doi.org/10.1101/2022.12.13.520232

Li, H., & Rauchfuss, T. B. (2002). Iron carbonyl sulfides, formaldehyde, and amines condense to give the proposed azadithiolate cofactor of the Fe-only hydrogenases. *Journal of the American Chemical Society*, *124*(5), 726–727.

Lloyd, D., Ralphs, J. R., & Harris, J. C. (2002). Giardia intestinalis, a eukaryote without hydrogenosomes, produces hydrogen. *Microbiology*, *148*(Pt 3), 727–733.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research*, *48*(D1), D265–D268.

Lyon, E. J., Georgakaki, I. P., Reibenspies, J. H., & Darensbourg, M. Y. (1999). Carbon Monoxide and Cyanide Ligands in a Classical Organometallic Complex Model for Fe-Only Hydrogenase. *Angewandte Chemie* , *38*(21), 3178–3180.

Ma, K., Schicho, R. N., Kelly, R. M., & Adams, M. W. (1993). Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(11), 5341–5344.

Martin, W., & Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. *Nature*, *392*(6671), 37–41.

Méheust, R., Castelle, C. J., Matheus Carnevali, P. B., Farag, I. F., He, C., Chen, L.-X., Amano, Y., Hug, L. A., & Banfield, J. F. (2020). Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog. *The ISME Journal*, *14*(12), 2907–2922.

Mistry, J., Bateman, A., & Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics*, *8*, 298.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419.

Moreira, D., & Lopez-Garcia, P. (1998). Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *Journal of Molecular Evolution*, *47*(5), 517–530.

Morita, R. Y. (1999). Is H(2) the Universal Energy Source for Long-Term Survival? *Microbial Ecology*, *38*(4), 307–320.

Mulder, D. W., Boyd, E. S., Sarma, R., Lange, R. K., Endrizzi, J. A., Broderick, J. B., & Peters, J. W. (2010). Stepwise [FeFe]-hydrogenase H-cluster assembly revealed in the structure of HydA(DeltaEFG). *Nature*, *465*(7295), 248–251.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), 555–560.

Nixon, J. E. J., Field, J., McArthur, A. G., Sogin, M. L., Yarlett, N., Loftus, B. J., & Samuelson, J. (2003). Iron-dependent hydrogenases of Entamoeba histolytica and Giardia lamblia: activity of the recombinant entamoebic enzyme and evidence for lateral gene transfer. *The Biological Bulletin*, *204*(1), 1–9.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834.

Nývltová, E., Šuták, R., Harant, K., Šedinová, M., Hrdy, I., Paces, J., Vlček, Č., & Tachezy, J. (2013). NIF-type iron-sulfur cluster assembly system is duplicated and distributed in the mitochondria and cytosol of Mastigamoeba balamuthi. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(18), 7371–7376.

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, *11*(12), 2864–2868.

Ortiz, M., Leung, P. M., Shelley, G., Jirapanjawat, T., Nauer, P. A., Van Goethem, M. W., Bay, S. K., Islam, Z. F., Jordaan, K., Vikram, S., Chown, S. L., Hogg, I. D., Makhalanyane, T. P., Grinter, R., Cowan, D. A., & Greening, C. (2021). Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45). https://doi.org/10.1073/pnas.2025322118

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy.

Nucleic Acids Research, *50*(D1), D785–D794.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* , *28*(11), 1420–1428.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Peters, J. W., Lanzilotta, W. N., Lemon, B. J., & Seefeldt, L. C. (1998). X-ray crystal structure of the Fe-only hydrogenase (CpI) from Clostridium pasteurianum to 1.8 angstrom resolution. *Science*, *282*(5395), 1853–1858.

Peters, J. W., Schut, G. J., Boyd, E. S., Mulder, D. W., Shepard, E. M., Broderick, J. B., King, P. W., & Adams, M. W. W. (2015). [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. *Biochimica et Biophysica Acta*, *1853*(6), 1350–1369.

Piché-Choquette, S., & Constant, P. (2019). Molecular Hydrogen, a Neglected Key Driver of Soil Biogeochemical Processes. *Applied and Environmental Microbiology*, *85*(6). https://doi.org/10.1128/AEM.02418-18

Pinske, C. (2019). Bioenergetic aspects of archaeal and bacterial hydrogen metabolism. *Advances in Microbial Physiology*, *74*, 487–514.

Probst, A. J., Castelle, C. J., Singh, A., Brown, C. T., Anantharaman, K., Sharon, I., Hug, L. A., Burstein, D., Emerson, J. B., Thomas, B. C., & Banfield, J. F. (2017). Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. *Environmental Microbiology*, *19*(2), 459–474.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(Database issue), D61–D65.

Rambo, I. M., Langwig, M. V., Leão, P., De Anda, V., & Baker, B. J. (2022). Genomes of six viruses that infect Asgard archaea from deep-sea sediments. *Nature Microbiology*, *7*(7), 953–961.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437.

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R. I., Xu, J., Rittmann, S. K.-M. R., Klingl, A., Pilhofer, M., & Schleper, C. (2023). Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature*, *613*(7943), 332–339.

Sato, T., Atomi, H., & Imanaka, T. (2007). Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science*, *315*(5814), 1003–1006.

Schleucher, J., Griesinger, C., Schwörer, B., & Thauer, R. K. (1994). H2-forming N5, N10-

methylenetetrahydromethanopterin dehydrogenase from Methanobacterium thermoautotrophicum catalyzes a stereoselective hydride transfer as determined by two-dimensional NMR spectroscopy. *Biochemistry*, *33*(13), 3986–3993.

Schmidt, M., Contakes, S. M., & Rauchfuss, T. B. (1999). First Generation Analogues of the Binuclear Site in the Fe-Only Hydrogenases: Fe2(μ-SR)2(CO)4(CN)22-. *Journal of the American Chemical Society*, *121*(41), 9736–9737.

Schuchmann, K., Chowdhury, N. P., & Müller, V. (2018). Complex Multimeric [FeFe] Hydrogenases: Biochemistry, Physiology and New Opportunities for the Hydrogen Economy. *Frontiers in Microbiology*, *9*, 2911.

Schuchmann, K., & Müller, V. (2012). A bacterial electron-bifurcating hydrogenase. *The Journal of Biological Chemistry*, *287*(37), 31165–31171.

Schut, G. J., & Adams, M. W. W. (2009). The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. *Journal of Bacteriology*, *191*(13), 4451–4457.

Schwartz, E., & Friedrich, B. (2006). The H2-Metabolizing Prokaryotes. In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, & E. Stackebrandt (Eds.), *The Prokaryotes: Volume 2: Ecophysiology and Biochemistry* (pp. 496–563). Springer New York.

Shima, S., Pilak, O., Vogt, S., Schick, M., Stagni, M. S., Meyer-Klaucke, W., Warkentin, E., Thauer, R. K., & Ermler, U. (2008). The crystal structure of [Fe]-hydrogenase reveals the geometry of the active site. *Science*, *321*(5888), 572–575.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843.

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science: A Publication of the Protein Society*, *27*(1), 135–145.

Smutná, T., Dohnálková, A., Sutak, R., Narayanasamy, R. K., Tachezy, J., & Hrdý, I. (2022). A cytosolic ferredoxin-independent hydrogenase possibly mediates hydrogen uptake in Trichomonas vaginalis. *Current Biology: CB*, *32*(1), 124–135.e5.

Søndergaard, D., Pedersen, C. N. S., & Greening, C. (2016). HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports*, *6*, 34212.

Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N., & Martin, W. F. (2016). Lokiarchaeon is hydrogen dependent. *Nature Microbiology*, *1*, 16034.

Spang, A., Caceres, E. F., & Ettema, T. J. G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, *357*(6351). https://doi.org/10.1126/science.aaf3883

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179.

Spang, A., Stairs, C. W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E. F., Greening, C., Baker, B. J., & Ettema, T. J. G. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nature Microbiology*, *4*(7), 1138–1148.

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1), 2542.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Valentin-Alvarado, L. E., Fakra, S. C., Probst, A. J., Giska, J. R., Jaffe, A. L., Oltrogge, L. M., West-Roberts, J., Rowland, J., Manga, M., Savage, D. F., Greening, C., Baker, B. J., & Banfield, J. F. (2024). Autotrophic biofilms sustained by deeply sourced groundwater host diverse bacteria implicated in sulfur and hydrogen metabolism. *Microbiome*, *12*(1), 15.

Vignais, P. M., & Billoud, B. (2007). Occurrence, classification, and biological function of hydrogenases: an overview. *Chemical Reviews*, *107*(10), 4206–4272.

Volbeda, A., Charon, M. H., Piras, C., Hatchikian, E. C., Frey, M., & Fontecilla-Camps, J. C. (1995). Crystal structure of the nickel-iron hydrogenase from Desulfovibrio gigas. *Nature*, *373*(6515), 580–587.

Wang, S., Huang, H., Kahnt, J., & Thauer, R. K. (2013). A reversible electron-bifurcating ferredoxin- and NAD-dependent [FeFe]-hydrogenase (HydABC) in Moorella thermoacetica. *Journal of Bacteriology*, *195*(6), 1267–1275.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., & Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, *1*(9), 16116.

Winkler, M., Esselborn, J., & Happe, T. (2013). Molecular basis of [FeFe]-hydrogenase function: an insight into the complex interplay between protein and catalytic cofactor. *Biochimica et Biophysica Acta*, *1827*(8-9), 974–985.

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* , *32*(4), 605–607.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* , *26*(13), 1608–1615.

Zaffaroni, R., Rauchfuss, T. B., Gray, D. L., De Gioia, L., & Zampella, G. (2012). Terminal vs bridging hydrides of diiron dithiolates: protonation of Fe2(dithiolate)(CO)2(PMe3)4. *Journal of the American Chemical Society*, *134*(46), 19260–19269.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358.

Zheng, Y., Kahnt, J., Kwon, I. H., Mackie, R. I., & Thauer, R. K. (2014). Hydrogen formation and its regulation in Ruminococcus albus: involvement of an electron-bifurcating [FeFe]-hydrogenase, of a non-electron-bifurcating [FeFe]-hydrogenase, and of a putative hydrogen-sensing [FeFe]-hydrogenase. *Journal of Bacteriology*, *196*(22), 3840–3852.

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., &

Anantharaman, K. (2020). METABOLIC: High-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. In *bioRxiv* (p. 761643). https://doi.org/10.1101/761643

Transitional section

While the roles of Asgard archaea in marine environments have been extensively studied, their potential contributions to soil ecosystems, particularly in the context of methane cycling in wetland soils, remain largely unexplored. In the following chapter, I reconstructed two complete genomes for soil-associated Atabeyarchaeia, a new Asgard lineage, and the first complete genome of Freyarchaeia. These high-quality genomes, obtained through manual curation and validated using nanopore long-read sequencing, provide a solid foundation for elucidating the metabolic capabilities and ecological roles of Asgard archaea in terrestrial ecosystems.

3. Asgard archaea modulate potential methanogenesis substrates in wetland soil

The following chapter is a modified version with permission of authors of the following published preprint: Luis E. Valentin-Alvarado, Kathryn E. Appler, Valerie De Anda, Marie C. Schoelmerich, Jacob West-Roberts, Veronika Kivenson, Alexander Crits-Christoph, Lynn Ly, Rohan Sachdeva, David F. Savage, Brett J. Baker and Jillian F. Banfield (2023). *In prep*. Asgard archaea modulate potential methanogenesis substrates in wetland soil

Abstract

The roles of Asgard archaea in eukaryogenesis and marine biogeochemical cycles are well studied, yet their contributions in soil ecosystems are unknown. Of particular interest are Asgard archaeal contributions to methane cycling in wetland soils. To investigate this, we reconstructed two complete genomes for soil-associated Atabeyarchaeia, a new Asgard lineage, and the first complete genome of Freyarchaeia, and defined their metabolism *in situ.* Metatranscriptomics highlights high expression of [NiFe]-hydrogenases, pyruvate oxidation and carbon fixation via the Wood-Ljungdahl pathway genes. Also highly expressed are genes encoding enzymes for amino acid metabolism, anaerobic aldehyde oxidation, hydrogen peroxide detoxification and glycerol and carbohydrate breakdown to acetate and formate. Overall, soil-associated Asgard archaea are predicted to be non-methanogenic acetogens, likely impacting reservoirs of substrates for methane production in terrestrial ecosystems.

N.B. *All main figures for this manuscript can be found below in section 3.6. All supplementary files (including figures and tables) can be found online with the published bioRxiv preprint.*

3.1 Introduction

Wetland soils are hotspots for methane production by methanogenic archaea. The extent of methane production depends in part on the availability of substrates for methanogenesis (e.g., formate, formaldehyde, methanol, acetate, hydrogen), compounds that are both produced and consumed by co-existing microbial community members. Among the groups of organisms that coexist with methanogens are Asgard archaea, of recent interest from the perspective of eukaryogenesis (Eme et al., 2023; Liu et al., 2021; Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017). To date, numerous lineages of Asgard archaea have been reported from anaerobic, sedimentary freshwater, marine, and hydrothermal environments (Cai et al., 2020, 2021; Eme et al., 2023; Farag Ibrahim F. et al., 2021; Imachi et al., 2020; Liu et al., 2018, 2021; Rodrigues-Oliveira et al., 2023; Seitz et al., 2016, 2019; Spang et al., 2015; Sun et al., 2021; Xie et al., 2022; Zaremba-Niedzwiedzka et al., 2017; Zhang et al., 2021). Predictions primarily from draft metagenome-assembled genomes (MAGs) indicate metabolic diversity and flexibility that may enable them to occupy these diverse ecological niches. It appears that Asgard archaea are not capable of methane production since they lack the key canonical methyl-coenzyme M reductase (MCR). Although a few complete genomes for Asgard from hydrothermal and geothermal environments have been reported (Imachi et al., 2020; Rodrigues-Oliveira et al., 2023; Tamarit et al., 2022; F. Wu et al., 2022), most metabolic analyses of Asgard archaea are limited by reliance on partial genomes. To date, no Asgard genomes from non-estuarine wetland soils have been reported. Thus, nothing is known about the ways in which Asgard archaea directly (via methane production) or indirectly (via metabolic interactions) impact methane cycling in wetlands.

To investigate the roles of Asgard archaea in carbon cycling in wetland soil, we reconstructed two complete genomes for a newly defined group, here named Atabeyarchaeia, and one complete genome for a group named Freyarchaeia. Freyarchaeia MAGs were originally reconstructed from Guaymas Basin, located in the Gulf of California, México (Xie et al., 2022), and from Jinze Hot Spring (Yunnan, China) (Eme et al., 2023). Subsequently, another group used the original data to recover similar genomes and referred to them as Jordarchaeia (Sun et al., 2022). Here, we retain the original nomenclature. The genomes for soil Asgard archaea were initially reconstructed by manual curation of Illumina short read assemblies and then validated using both Nanopore and PacBio long reads. These fully curated genomes enabled us to perform comprehensive metabolic analyses, without the risks associated with reliance on draft genomes, and provided context for metatranscriptomic measurements of their *in situ* activity. Our integrated analysis of gene expression and metabolic predictions revealed roles for Atabeyarchaeia and Freyarchaeia in the production and consumption of carbon compounds that can serve as substrates for methanogenesis by coexisting methanogenic archaea.

3.2 Materials and Methods

Sample acquisition, nucleic acid extraction and sequencing

We collected soil cores from a seasonal vernal pool in Lake County, California, in October 2018, October 2019, November 2020 and October 2021. Samples were frozen in the field using dry ice, and kept at -80 C until extraction. The Qiagen PowerSoil Max DNA extraction kit was used to extract DNA from 5-10 g of soil, and the Qiagen AllPrep DNA/RNA extraction kit was used to extract RNA from 2 g of soil. Samples were sequenced by the QB3 sequencing facility at the University of California, Berkeley on a NovaSeq 6000. Read lengths for the 2018 DNA samples and the RNA samples were 2x150 bp and 2x250 bp for the 2019-2021 DNA samples. A sequencing depth of 10 Gb was targeted for each of 2018, 2020, and 2021 samples, and 20 Gbp for each of the 2019 samples. A subset of deep soil samples from 2021 were sequenced using PacBio and Oxford Nanopore technologies. For Oxford Nanopore sequencing, the Ligation Sequencing Kit (LSK114) was used to prepare native DNA libraries and the Qiagen Repli-G mini kit and LSK114 were used to prepare amplified DNA libraries. Samples were sequenced on FLO-PRO114M flowcells on the PromethION for 72 hours.

Illumina metagenomic assembly, binning and annotation

Metagenomic sequencing reads were trimmed for adapter sequences and quality using sickle (*Sickle: Windowed Adaptive Trimming for Fastq Files Using Quality*, n.d.). The filtered metagenomic reads from 2018 were assembled using the IDBA_UD assembler (Peng et al., 2012) . The reads from 2019-2020 using metaSPAdes(Nurk et al., 2017) Contigs greater than 2.5 Kb were retained and sequencing reads from all samples were cross-mapped against each resulting assembly using Bowtie2 (Langmead and Salzberg 2012). The resulting differential coverage profiles were filtered at a 95% read identity cutoff and then used for genome binning with MetaBAT2(Kang et al., 2019), VAMB (Nissen et al. 2021) and MaxBin2 (Wu et al. 2016). The resulting genome bins were assessed for completeness and contamination using CheckM2 (Parks et al., 2015) and were manually curated using taxonomic profiling with GGKBase (www.ggkbase.berkeley.edu). The first iteration of taxonomy was assigned to genome bins with GTDB-Tk v2.3.0 (Chaumeil et al., 2019) and further validated with phylogenetic trees of single-copy marker genes. We also downloaded all the publicly available Asgard genomes from BV-BRC (www.bv-brc.org) and used CheckM2 to estimate genome completeness.

PacBio metagenomic sequencing and assembly

Samples from September 9, 2021 from 140 cm and 75 cm were sequenced using a Sequel II to generate PacBio HiFi reads. Reads were quality trimmed using BBDuk (bbduk.sh minavgquality=20 qtrim=rl trimq=20) (Bushnell, 2014) and assembled with hifiasm-meta (Feng et al., 2022).

Oxford Nanopore metagenomic assembly

Reads were basecalled with Guppy using the dna_r10.4.1_e8.2_sup@v3.5.1 model. Reads were filtered for an average quality >10 and a minimum length of 1kb using fastp (v0.23.2). Adapters were trimmed using porechop (v0.2.4). Branching artifacts caused by multiple displacement amplification were removed by aligning amplified reads to themselves with mappy (v2.24); Reads with non-diagonal self-alignment were removed. Both amplified and native reads were jointly assembled with metaFlye (v2.9) (Kolmogorov et al., 2020), long-read polished with medaka consensus (v1.7.1), and short-read polished with Hapo-G (v1.3.1).

Manual genome curation of short-reads genomes and validation using long-reads

The manually curated genomes were de novo reconstructed from high-quality Illumina metagenomic data as described previously (Chen et al. 2020). From soil samples taken at various depths, we recovered draft illumina based MAGs corresponding to Atabeyarchaeia-1, Atabeyarchaeia-2 and Freyarchaeia. The curation process involved the identification and removal of obvious chimeric regions, which were indicated by abrupt changes in GC content or by insufficient Illumina read mapping support. We also corrected sequences in regions with imperfect read alignment, allowing no single nucleotide polymorphisms (SNPs), by mapping reads at a reduced stringency threshold (allowing for up to 3% SNPs). This was followed by manual curation of the consensus sequence, including insertion, deletion, or substitution of individual base pairs. The extension of contig ends was conducted using unplaced Illumina reads. High read coverage was interpreted as indicative of the genomic termini. A genome was deemed complete when it displayed uninterrupted support from Illumina reads. The final assessment of genome completeness was performed by examining the cumulative GC skew and ensuring alignment with known complete genomes from related taxa. The Average Amino Acids of the new genomes was performed using AAI: Average Amino acid Identity calculator tool (http://enve-omics.ce.gatech.edu/aai/) and using compareM (v.0.0.23) with the 'aai_wf' at default settings (https://github.com/dparks1134/CompareM). Replichores of complete genomes were predicted according to the GC skew and cumulative GC skew calculated by the iRep package (gc_skew.py).

Metabolic pathway reconstruction of complete genomes and phylogenetic analysis of key functional genes

We utilized Prodigal (v2.6.3) to predict the genes for Atabeyarchaeia and Freyarchaeia genome by this study (Hyatt et al., 2010). A suite of databases, including KofamKOALA (v1.3.0) (Aramaki et al., 2020) and InterProScan (5.50-84.0) (Paysan-Lafosse et al., 2023) were combined to begin annotation. Complete genomes were also annotated with HydDB(Søndergaard et al., 2016), METABOLIC (v4.0) (Zhou et al., 2022), PROKKA (v1.14.6), DRAM (Shaffer et al., 2020) and MEBS (v2.0). Our methods also incorporated a microbial genome annotation workflow as previously described in Undinarchaeota study (Dombrowski et al., 2020).

For each gene/protein of interest, references were compiled by 1) BLASTing the corresponding gene against the NCBI nr database and their top 50 hits clustered by CD-HIT using a 90% similarity threshold, 2) obtaining sequences from the high quality manually annotated UniProtKB/Swiss-Prot/InterPro/PFAM (reviewed) (*catalases*), or 3) using previously published alignment and creating hmm models to extract Atabeyarchaeia and Freyarchaeia sequences (*AOR, catalases,* and *MttB-like homologs*). The final set genes/proteins were aligned using MAFFT v.7.407 (*AOR* and *catalases*)/v7.490 (*ADH* and *MttB-like homologs*), trimmed with trimAl v1.4.rev15 (*AOR, ADH*, *catalases,* and *MttB-like homologs*), and a phylogenetic tree was inferred using IQTREE v.1.6.6 (*AOR* and *catalases*)/v2.0.7 (*ADH* and *MttB-like homologs*) using automatic model selection. More detailed descriptions for each phylogeny, including model selection are discussed in the supplemental figure captions.

Identification of [NiFe]-hydrogenase and phylogenetic analyses

We extracted [NiFe]-hydrogenase sequences from complete genomes as well as publicly available Asgardarchaeota genomes from NCBI and BCWT using custom Hidden Markov Models (HMMs). To confirm the accuracy of the candidate [NiFe]-hydrogenase sequences, we checked the identified sequences by looking at nearby genes and ensuring the presence of essential hydrogenase accessory genes and confirmed the preliminary classification using HydDB(Søndergaard et al., 2016). We combined our [NiFe]-hydrogenase sequences with those from the HydDB database. Then, we aligned all the sequences using MAFFT (parameters:--localpair --maxiterate 1000). After alignment, we looked for the N-terminal and C-terminal CxxC conserved motifs. We then used TrimAl v (parameters: -gt 0.5) to clean up the alignment, removing parts where more than 50% of the sequences had gaps. We used IQ-TREE version 1.6.12 with the best-fit model according to Bayesian information criterion (BIC) and ultrafast 1000 bootstrap method to estimate support values for the tree branches. The output tree was visualized using ItoL and the subgroups were classified based on known HYDB references (Søndergaard et al., 2016).

Identification of Selenocysteine machinery and Selenoproteins, tRNA and intron predictions

The Sec machinery (pstK, EFsec, SecS, and SPS) was identified using HMMER v3.3.1 (Eddy 2011) via hmmscan against the TIGRFAM v15.0 database (Li et al. 2021) and also confirmed via Selenoprofiles v4.4.8 (Santesmasses et al., 2018), using the -p machinery option. Selenoprofiles v4.4.8 and the Seblastian- SECIS3 software (Mariotti et al. 2013) was accessed to identify potential selenoproteins, which were manually curated via alignment to known selenoproteins as previously reported (Mariotti et al., 2016). SECIsearch3 was used to identify SECIS for selenoproteins.

For identification of predicted pre-tRNA, we used tRNAscan-SE v.2.0.12 (default settings and -A for archaea model) (Chan et al. 2021). The R2DT software (Sweeney et al. 2021) was used to predict and visualize tRNA secondary structure, which also led to identification of additional introns that were not detected by tRNAscan. We identified Sec tRNA with the Secmarker v0.4 webserver (Santesmasses et al. 2017) with a minimum Infernal score of 35.

Metatranscriptome sequencing and analysis

A subset of the deep SRVP soil samples (60, 70, 80 and 100 cm) were used for metatranscriptomics. Asgard genomes were first dereplicated using derep v 3.4 (Olm et al., 2017), resulting in a set of representative genomes. A Bowtie2 index was constructed from the dereplicated genomes to expedite the sequence alignment process. Post alignment, a custom-built Python script (Crits-Christoph, n.d.) was utilized to scrutinize the paired reads from the alignment (BAM) files and deliver statistical insights on gene expression. This script initiates by creating a Pysam 'AlignmentFile' object, enabling a systematic traversal through the '.bam' files by parsing the complex binary alignments. It then proceeds to iterate over each gene listed in a predefined set. For each gene, the script tallies the reads that satisfy user-defined thresholds for Average Nucleotide Identity (ANI) and Mapping Quality (MAPQ). These thresholds, designed to uphold a high level of alignment quality and sequence identity, are critical in enhancing the reliability of the data. Upon the completion of this process, the script outputs the filename, the gene name, and the count of reads that conform to the defined parameters. This pipeline permits an in-depth investigation of RNA hits within the dereplicated genomes, thereby illuminating gene expression profiles across diverse environmental scenarios.

To quantify the number of reads per gene, we further utilized a Python script (filter_counts.py). This script operates based on gene predictions produced by the Prodigal software for the genomes of interest. With the required gene predictions secured through a rerun of Prodigal, the script was set in motion on the BAM files. The command executed was as follows: python filter_counts.py -q 10 -m 0.97 derep_genomes.faa. This command sets the minimum MAPQ score (-q 10) and the minimum ANI (-m 0.97) to ensure high-quality alignments and a high degree of sequence identity. This adherence to stringent criteria bolsters the reliability and robustness of our analytical process.

Mapped mRNA reads were standardized by the calculation reads per kilobase of transcript per million reads mapped (RPKM), which allows for the direct comparison of the level of transcription of all genes in the de novo assembly (Table S4). RPKM is calculated in Rockhopper as: RPKM=((number of mapped reads/length of transcript (gene) in kilobase)/million mapped reads)) (McClure et al., 2013).

Abundance and distribution of archaea

The dereplicated Atabeyarchaeia, Freyarchaeia and other archaeal genomes present in the 36 metagenome samples were mapped to all the metagenome reads independently using BBMap (Bushnell, 2014) (parameters: nodisk=t pigz=t unpigz=t ambiguous=random). We set the minimum identity for mapping to 0.9 and handled ambiguous mappings randomly. The percentage of relative abundance of each genome, and the unmapped read percentage values were calculated using coverM (parameters: -m relative_abundance --min-read-percent-identity 95).

Phylogenetic analyses of Asgard genomes

We used three different sets of markers for the phylogenetic analysis. 47 arCOGs (Archaeal cluster of orthologous genes), a subset of previously described marker set (Baker et al., 2016), were extracted with hmmsearch (v3.1b2) from the 9 MAGs in this study, 303 additional Asgard MAGs, and 36 TACK publicly available MAGs. Korarchaeales MAGs were excluded from the analysis due to the hyperthermal-sensitive marker genes (Eme et al., 2023). We chose to exclude arCOG01183 from the previous set as it was found in less than 50% of the MAGs. mafft (v7.310) and BMGE (v1.12) were used to align and trim the concatenated alignments. Maximum likelihood phylogenetic trees were generated using IQ-TREE (v1.6.1) to test various models and topologies to obtain bootstrap values and LG+F+R10 model was selected (Figure 3.1C). Additionally, we compared this phylogeny with a set of 37 single-copy marker genes from Phylosift (v1.0.1) (Darling et al., 2014) (Darling et al., 2014). The sequences were manually accessed, aligned with mafft auto (v7.490), trimmed with BMGE (v2.0), and a maximum likelihood phylogeny generated with IQ-TREE (v2.0.7), using the LG+F+R10 model (Figure S24). All marker sets resulted in the same phylogeny for the 9 added MAGs. 16S ribosomal proteins were extracted from 439 Asgard and 39 TACK MAGs with barrnap 0.9, using options --kingdom arc --lencutoff 0.2 --reject 0.3 --evalue 1e-05. Sequences manually curated, aligned with mafft auto (v7.490) and masked with Geneious Prime, and ran with IQ-Tree (v2.0.7). 16S phylogeny separates Atabeyarchaeia MAGs from other Asgard clades (Figure S2). Amino acid identity (AAI) was determined with CompareM (v0.0.23) to distinguish taxonomic level. We assigned numbers from 1-9 as candidate groups within this new Asgardarchaeota class.

Identification and analysis of genomic ESPs

PSI-BLAST was used to query the Asgard proteomes against the AsCOGs (Asgard database) and arCOGs databases (Liu et al., 2021). Hits were only considered if the absolute difference between the subject and query sequence was less than 75% of the length of the query sequence. Hits were then dereplicated by taking the best hit for each query.

Protein structure prediction using AlphaFold-multimer and ColabFold

Structural predictions for group 4 [NiFe]-hydrogenases, CoxLMS, MtrA, and MtrAH fusion proteins were performed using ColabFold, incorporating AlphaFold-multimer capabilities (Mirdita et al. 2022). These predicted structures were then visualized and aligned using ChimeraX (Meng et al., 2023), with corroborating reference structures obtained from the Protein Data Bank (Berman et al., 2000). This structural modeling was integral in reinforcing our phylogenetic and metabolic inference.

Nomenclature, etymology, and proposal of type material

Based on the findings of our research, we propose a new taxonomic nomenclature for the identified Asgard archaea organism. The chosen nomenclature encapsulates various facets that resonate with the organism's nature and its ecological habitat. The term 'Atabey' is a homage to the Mother Goddess, often referred to as the Earth Mother in the Taíno mythology, symbolizing fertility, abundance, and the nurturing aspects of nature. We suggest the taxonomic hierarchy of class 'Ca. Atabeyarchaeia', order 'Ca. Atabeyarchaeales', family 'Ca. Atabeyarchaeaceae', and genus 'Ca. Atabeyarchaeum'. We submit the complete genomes "Atabeyarchaeia-1" and "Atabeyaarcaeia-2" as well 6 near complete genomes as type material for this new group.

3.3 Results

Complete genomes and phylogenetic placement of Asgard archaea from wetland soil

We analyzed Illumina metagenomic data from samples collected from 20 cm to 175 cm depth in the soil of a wetland located in Lake County, California, USA. We previously reported megaphages (Al-Shayeb et al. 2020) and *Methanoperedens* archaea and their 1 Mb-scale "Borg" extrachromosomal elements from this site (Al-Shayeb et al. 2022). From the metagenomic analyses conducted at this site, we determined that archaea account for >45% of the total community below a depth of 60 cm. Archaeal groups detected include members of the Asgardarchaeota, Bathyarchaeia, Methanosarcinia, Nitrososphaeria, Thermoplasmata, Micrarchaeia, Diapherotrites, Aenigmatarchaeia, Methanomicrobia, Aenigmarchaeia, Nanoarchaeia, Hadarchaeia and Methanomethylicia (Figure 3.1A-B).

From 60, 80, and 100 cm deep wetland soil, we recovered four draft Asgard genomes, three of which were manually curated to completion using methods described previously (Chen et al. 2020). Taxonomic classification using the Silva DB placed the 3,576,204 bp genome as Freyachaeia. 16S rRNA gene sequence analysis showed the two other complete genomes were distinct from Freyarchaeia (16S rRNA genes are <75% identical), thus representing organisms from a separate, new lineage. These genomes are 2,808,651 and 2,756,679 bp in length (Table S1) with an average amino acid identity (AAI) of ~70% (Table S2).

Phylogenetic analyses using several sets of marker genes ("see materials and methods") placed our two novel complete genomes in a monophyletic group within the Asgard clade as a sister group to Freyarchaeaia (Figure 3.1C). We performed phylogenetic analyses using concatenated marker sets of 47 arCOG and 15 ribosomal protein (RP15) gene cluster (Figure S1), as well as 16S rRNA (Figure S2). The new genomes share only 40-45% AAI when compared to other Asgard genomes, consistent with their assignment to a new phylum. Although our analyses provide evidence for distinction at the phylum level, we chose to adhere to the Genome Taxonomy Database (GTDB) for standardized microbial genome nomenclature (Table S3). Here, we propose the name *Candidatus* "Atabeyarchaeia" for this new group, where 'Atabey' is a goddess in of Taíno Puerto Rican mythology. Atabeyarchaeia is represented by the complete Atabeyarchaeia group 1 (Atabeya-1) and group 2 (Atabeya-2) genomes. Included in this group are 2 MAGs from a highly fragmented, partial Asgard Lake Cootharaba Group (ALCG) draft genome (Sun et al. 2021). The cumulative GC skew of the Freyarchaeia and Atabeyarchaeia genomes is consistent with bidirectional replication. This style of replication is typical of bacterial genomes but has not been widely reported in Archaea, and has never been described in the Asgard group (Figure 3.1D and Figure S3).

Unexpectedly, we found that 92% to 95% of tRNA genes from all three genomes contain at least one intron. This contrasts with the general estimate that 15% of archaeal tRNA harbor introns (Marck & Grosjean, 2003), and with Thermoproteales (another order of archaea), where 70% of the tRNAs contain introns (Sugahara et al., 2008). In total, there are 228 tRNA introns across the three new Asgard genomes (Table S4). Unlike most archaeal tRNA introns that occur in the anticodon loop at position 37 / 38 (Tocchini-Valentini et al. 2011; Yoshihisa 2014), Atabeyarchaeia and Freyarchaeia introns often occur at non-canonical positions, and over half of their tRNA genes have multiple introns (Table S4).

Subsequently, we acquired and independently assembled Oxford Nanopore and PacBio long-reads from a subset of the samples to generate three circularized genomes that validate the overall topology of all three curated Illumina read-based genomes (Figure S4, Table S1). These complete genomes allowed us to genomically describe two Atabeya-2 strain variants from 100 cm and 175 cm depth soil. In addition, we used Illumina reads to curate a draft Nanopore genome for another Atabeyarchaeia species, Atabeya-3, from 75 cm and 175 cm depth soil (Figure S5). The Atabeyarchaeia-3 genome is most closely related to the Asgard Lake Cootharaba Group (ALCG) fragments (Sun et al. 2022). To further solidify the phylogenetic position of Atabeyarchaiea, we included the Atabeyarchaeia-3 genome and another draft genome (Atabeyarchaeia-4) from Illumina reads in the phylogenetic analysis.

Using the Asgard clusters of orthologous genes (AsCOGS) database and functional classification, we identified eukaryotic signature proteins (ESPs) in the complete and public genomes of Atabeyarchaeia and Freyarchaeia (Zaremba-Niedzwiedzka et al. 2017; Liu et al. 2021). Atabeyarchaeia and Freyarchaeia genomes had the highest percentage of hits for 'Intracellular trafficking, secretion, and vesicular transport' (U) among the AsCOG functional classes, accounting for 84.3% of the hits to the database. Within this class, we identified key protein domains such as Adaptin, ESCRT-I-III complexes, Gelsolin family protein, Longin domain, Rab-like GTPase, Ras family GTPase, and Roadblock/LC7 domain (Table S5, Figure S6). The 'Post Translational modification, protein turnover, and chaperones' category (O) followed with a count of 101 (15.8%), highlighting domains like Ubiquitin, Jab1/MPN domain-containing protein, and the RING finger domain. The presence of ESPs in the newly described Atabeyarchaeia lineage and their presence in Freyarchaeia aligns with previous findings for Asgardarchaeota (Eme et al. 2023; Spang et al. 2015; Liu et al. 2021).

Expression of energy conservation pathways constrain key metabolisms in situ

We analyzed the metabolic potential of the three complete genomes and investigated their activity *in situ* through metatranscriptomics of soil samples ("see materials and methods", Figure 3.2, Table S6, Table S7). The metatranscriptomic data indicate high expression of genes involved in key energy conservation pathways (Figure 3.3A). Most highly transcribed genes are soluble heterodisulfide reductase (HdrABC), [NiFe] hydrogenases (groups 3 and 4), ATP synthase, numerous aldehyde ferredoxin oxidoreductase genes, genes for phosphoenolpyruvate (PEP) and pyruvate metabolism, and carbon monoxide dehydrogenase/acetyl CoA synthetase (CODH/ACS). Notably, the Hdr, the group 3 and group 4 hydrogenase (including up to eight NADH-quinone oxidoreductase subunits, e.g., Nuo-like) as well as the ATP synthase are co-encoded in a syntenic block in all of the genomes (Figure 3.4A). Phylogenetic analysis of the large subunit of group 4 [NiFe]-hydrogenases suggests they are closely related to those of Odinarchaeia, Heimdallarchaia, and Hermodarchaeia (Figure 3.4B, Table S8). However, the exact function of this unclassified Asgard group has not been validated biochemically (Spang et al., 2019). One clue relies on the identification of eight genes homologous to the hydrophobic subunits of complex I NuoL, M, and N (*E. coli* nomenclature) and Mrp-type Na+/H+ antiporters. Thus, these Asgard archaea may mediate Na+/H+ translocation coupled to energy generation via ATP synthase (Yu et al. 2018; Marreiros et al. 2013; Efremov and Sazanov 2012).

We employed AlphaFold2 to model the hydrogenase and associated complex I-like modules. Overall, the predicted structure has a cytosolic and membrane-associated portion (Figure 3.4C). The cytosolic portion aligned with the respiratory membrane-bound hydrogenase (MBH) from *Pyrococcus furiosus* (Yu et al. 2018) with high confidence (Figure 3.4D). When superimposed, the calculated structures of the membrane-associated hydrophobic L, M, K, and S chains aligned to bacterial complex I. In the canonical complex I (Efremov and Sazanov 2011; Baradaran et al. 2013), Chain L, Nqo12, as well as M, N, and K translocate proteins (Nakamaru-Ogiso et al. 2010; Baradaran et al. 2013), a process that is facilitated by an arm, helix HL that is part of chain L. This helix HL is also present in the L-like subunit of the Asgard complexes (Figure 3.4E). The helix HL, and the antiporter subunits located between chain L and the subunit that connects to the cytosolic hydrogenase portion, are absent in all characterized respiratory membrane-bound hydrogenases (Figure 3.4E, Figure S7).

The Group 3c cofactor-coupled bidirectional [NiFe] hydrogenase (Figure S8A) in combination with HdrABC suggests the capability to bifurcate electrons from H_2 to ferredoxin and an unidentified heterodisulfide compound. This capacity has been observed in methanogenic archaea via the MvhADG–HdrABC system (Buckel and Thauer 2013; Sousa et al. 2016). Atabeyarchaeia genomes encode two independent gene clusters of the Group 3b NADP-coupled [NiFe] hydrogenases (Figure S8B). Their presence suggests the capacity to maintain redox equilibrium and, potentially, grow lithoautotrophically by using H_2 as an electron donor, as suggested for other Asgardarchaeota members (Spang et al. 2019; Xie et al. 2022; Zhang et al. 2021).

Atabeyarchaeia and Freyarchaeia encode both the tetrahydromethanopterin (H_4MPT) methyl branch and the carbonyl branch of the Wood–Ljungdahl pathway (WLP) (Figure S9). This reversible pathway can be used to reduce CO_2 to acetyl coenzyme A (acetyl CoA), which can be further converted to acetate. This last conversion can lead to energy conservation in both Asgard lineages via substrate-level phosphorylation when mediated by acetate-CoA ligase (see below).

We confirmed the expression of almost all of the genes of the methyl and carbonyl branches, including the acetate-CoA ligase, in all complete genomes. When H_2 is present in the ecosystem, these archaea could use the WLP for the reduction of CO_2 or formate and thereby conserving energy. Alternatively, they could use the WLP in reverse to oxidize acetate. In both scenarios, the expression of energy-converting hydrogenases and the ATP synthases suggest a potential role in energy conservation. This involvement may include coupling exergonic electron transfer to establish an ion gradient that fuels the ATP synthase for ATP generation. The metabolic inferences along with the transcriptional data including the expression of *por* genes in all three Asgard genomes, indicates a reliance on an archaeal version of the WLP to perform acetogenesis (Sousa et al. 2016; Orsi et al. 2020). This acetogenic lifestyle appears to involve energy conservation through a hydrogenase-dependent chemiosmotic mechanism similar to that observed in some acetogenic bacteria (Schoelmerich and Müller 2019).

Potential for non-methanogenic methylotrophic life-style and carboxydotrophy

Despite the absence of the MCR complex, Freyarchaeia genomes have all the necessary genes to synthesize coenzyme-M from sulfopyruvate via the ComABC pathway similar to methanogens (Graham & White, 2002; White, 1985). Most methanogens conserve energy via the Na+-translocating MtrA-H complex, which is encoded by an eight-gene cluster (Gottschalk and Thauer 2001). Although Atabeyarchaeia and Freyarchaeia do not have the genes for the full complex, Atabeyarchaeia-1 has two copies of the CH_3-H_4MPT-dependent methyltransferase subunit A-like (MtrA) and both Freyarchaeia and Atabeyarchaeia also encode the CH_3-H_4MPT-dependent methyltransferase subunit H (MtrH), along with a phylogenetically distinct fused polypeptide of MtrA-like and MtrH (Figure 3.5A). Under the conditions prevalent at the time of sampling, the *mtr* genes were only weakly expressed (Table S6, S8). While the biochemical activity of these divergent non-methanogen-associated MtrA-like and MtrH-like enzymes remain unclear, our phylogenetic analyses suggest they are phylogenetically related to methanogenic MtrA, MtrH, and MtrAH sequences. This suggests their potential role in converting CH_3-H_4MPT to H_4MPT, transferring a methyl group to an acceptor –possibly coenzyme-M, which can be produced by Freyarchaeia–. As they lack the MCR complex, the subsequent fate of the methyl group remains uncertain.

Although Atabeyarchaeia and Freyarchaeia genomes do not encode MrtE, we identified genes associated with methyltransferase systems encoded in close proximity to the MtrH gene. Specifically, the genomes encode trimethylamine methyltransferase (MttB-like, COG5598 superfamily), undefined corrinoid protein (MtbC-like), and putative glycine cleavage system H (gcvH) (Table S9). Both Atabeyarchaeia and Freyarchaeia genomes encode trimethylamine methyltransferase MttB (COG5598). Phylogenetic analysis suggests that MttB (Figure S10) and MtbC (Figure S11) belong to a previously uncharacterized group of methyltransferases, similar to those found in Njordarchaeales, Helarchaeales, Odinarchaeia and TACK members, including Brockarchaeia and Thermoproteota. In methanogens that encode *mttB*, this gene has an amber codon encoding the amino acid pyrrolysine in the active site (Hao et al. 2002; Li et al. 2023). The archaea from this study do not encode pyrrolysine, suggesting Freyarchaeia and Atabeyarchaeia encode a non-pyrrolysine MttB homolog, likely a quaternary amine (QA) dependent methyltransferase (Ticak et al. 2014). Only a fraction of QA methyltransferase substrates have been identified, and these include glycine betaine, proline betaine, carnitine, and butyrobetaine (Ticak et al. 2014; Picking et al. 2019; Kountz et al. 2020; Ellenbogen et al. 2021). The methyl group from the QA may be transferred to THF or H_4MPT branches of the WLP, akin to the mechanisms described in archaea with the capacity for non-methanogenic anaerobic methylotrophy, including Freyarchaeia (Jordarchaeia), Sifarchaeia, Brockarchaeia, and Culexarchaeia (Sun et al. 2021; De Anda et al. 2021; Farag Ibrahim F. et al. 2021; Kohtz et al. 2022). Consumption of QA compounds may reduce the pool of potential substrates for methanogenic methane production.

We identified genes in the Freyarchaeia genome that potentially encode an aerobic carbon-monoxide dehydrogenase complex (CoxLMS) and associated cofactors. Phylogenetic analysis places the putative CoxL in a monophyletic group with other archaea including Thermoplasmatales, Marsarchaeota, and Culexarchaeia (Figure S12). The gene cassette arrangement suggests these archaea may possess the ability to use carbon monoxide as a growth substrate (carboxydotrophy). However, analysis of the protein sequence reveals that the putative large-subunit aerobic CO dehydrogenases (CoxL) are missing the characteristic VAYRCSFR motif, which is critical for CO binding in the form I Cox proteins (Dobbek et al. 2002; Cordero et al. 2019). Nevertheless, the modeled protein structure, along with the operon organization of the *cox* genes, points to a novel type of Cox system in archaea (Figure S13). Alternatively, it is possible that this complex enables the utilization of alternative substrates, such as aldehydes or purines, as a member of the aldehyde oxidase superfamily (Kohtz et al. 2022; Cordero et al. 2019; Beam et al. 2016).

Carbon compound metabolic pathways

There are indications that Freyarchaeia and Atabeyarchaeia display distinct metabolic preferences for various soil carbon compounds (Figure 3.2; Figure S13). Freyarchaeia exhibit a genetic repertoire to break down various extracellular lignin-derived compounds including 5-carboxyvanillate. Other substrates that we predict can be metabolized by Freyarchaeia carbohydrate-active enzymes include hemicellulose (C5), cellobiose, maltose, and cellulose (C12). We predict that cellodextrin (C18) compounds can be converted to glucose via beta-glucosidase (BglBX). The findings implicate Freyarchaeia in the metabolism of plant-derived soil carbon compounds. Glucose, resulting from the degradation of complex carbohydrates, as well as ribulose and other carbon substrates, likely enters the modified Embden-Meyerhof-Parnas (EMP) pathway, yet the genes of this EMP pathway genes are only weakly expressed (Figure S14, Table S6, S8**)**. Additionally, Freyarchaeia encode and express an array of genes for the uptake of carbohydrates including major facilitator superfamily sugar transporters and ABC-sugar transporters suggesting an active role in efficiently assimilating diverse carbon substrates from soil environments Atabeyarchaeia also harbor genes of the EMP glycolytic pathway, producing ATP through the conversion of acetyl-CoA to acetate (Figure S13). Unlike Freyarchaeia which likely feed glucose into the EMP pathway, the entry point for Atabeyarchaeia to the EMP pathway appears to be fructose 6-phosphate (F6P). This is relatively uncommon for Asgard archaea but is reminiscent of the pathway in Helarchaeales (Seitz et al. 2019), an order of Lokiarchaeia. We identified Atabeyarchaeia transcripts for all but one of the genes for the steps from G6P to acetate (Table S8).

Atabeyarchaeia and Freyarchaeia utilize different enzymes to produce pyruvate. Atabeyarchaeia encode the oxygen-sensitive reversible enzyme, pyruvate:phosphate dikinase (PpdK); whereas Freyarchaeia encodes unidirectional pyruvate water dikinase/phosphoenolpyruvate synthase (PpS) and pyruvate kinase (Pk), producing phosphoenolpyruvate and pyruvate (Bräsen et al. 2014), respectively. Pyruvate generated via EMP pathway can be then converted to acetyl-CoA by pyruvate:ferredoxin oxidoreductase (PorABCDG) complex using a low-potential electron carrier such as a ferredoxin as the electron donor. Alternatively, acetyl-CoA can also be generated via pyruvate formate-lyase (pflD) generating formate as a byproduct. The final step involves the conversion of acetyl-CoA to acetate via acetate-CoA ligase (ADP-forming) producing ATP via substrate level phosphorylation- a crucial energy conserving step during fermentation of carbon compounds in both lineages.

Lacking the ability to phosphorylate C6 carbon sources, Atabeyarchaeia converts ribulose-5-phosphate (C5) and fixes formaldehyde (C1) into hexulose-6-phosphate (H6P) via the ribulose monophosphate (RuMP) and non-oxidative pentose phosphate (NO-PPP) pathways (Figure 3.2, Figure S14). The Atabeyarchaeia RuMP pathway bifunctional enzymes (HPS-PH and Fae-HPS) are common in archaea and similar to methylotrophic bacterial homologs (Kato et al. 2006). The RuMP pathway in these Asgard archaea can modulate the formaldehyde availability, a byproduct of methanol oxidation, microbial organic matter decomposition, and combustion. High expression of aldehyde-ferredoxin oxidoreductases (AOR) genes suggests another mechanism for the interconversion of organic acids to aldehydes. For example, aldehyde detoxification (e.g., formaldehyde to formate) and source of acetate from acetaldehyde Atabeyarchaeia-1, Atabeyarchaeia-2, and Freyarchaeia encode multiple AOR gene copies 5, 6, and 8 respectively. Phylogenetic analyses (Figure S15) suggest that both Asgard lineages encode AOR genes related to the FOR family that oxidize C1-C3 aldehydes or aliphatic and aromatic aldehydes (e.g. formaldehyde or glyceraldehyde) (Arndt et al., 2019; Bevers et al., 2005; Mukund & Adams, 1995). Furthermore, Freyarchaeia also encodes a tungsten-based AOR-type enzyme (XOR family) found in cellulolytic anaerobes with undefined substrate specificity (Scott et al. 2015) (Figure S15). Of the classified AORs, only one gene is expressed in Atabeyarchaeia-2 (Figure 3.3). Yet, some of the unclassified AOR genes are among the most highly expressed genes in the Atabeyarchaeia genomes. Despite the lack of biochemical characterization for most AOR families, these observations suggest a key role of multiple aldehydes in the generation of reducing power in the form of reduced ferredoxin.

Similar to other Asgard archaea (Seitz et al. 2019; Zhang et al. 2021; Spang et al. 2019), Atabeyarchaeia and Freyarchaeia encode genes for the large subunit of type IV and methanogenic type III Ribulose 1,5-bisphosphate carboxylase (RbcL) (Figure S16) a key enzyme in the partial nucleotide salvage pathway. This pathway facilitates the conversion of adenosine monophosphate (AMP) to 3-phosphoglycerate (3-PG), potentially leading to further metabolism into acetyl-CoA (Tabita et al. 2008).

Anaerobic glycerol (C3) metabolism by Atabeyarchaeia and Freyarchaeia is predicted based on the presence of glycerol kinase (GlpK), which forms glycerol-3-phosphate (3PG) from glycerol. 3PG (along with F6P) can be broken down via the EMP pathway or 3PG can be converted to dihydroxyacetone phosphate (DHAP) via GlpABC. DHAP can also serve as a precursor for sn-glycerol-1-phosphate (G1P), the backbone of archaeal phospholipids. Freyarchaeia have an extra GlpABC operon, the GlpA subunit of which clusters phylogenetically with GlpA of Halobacteriales, the only known archaeal group capable of glycerol assimilation (Figure S17).

All three genomes have a partial TCA cycle similar to other anaerobic archaeal groups such as methanogens (Lang et al. 2015). They encode succinate dehydrogenase, succinyl-CoA synthetase, 2-oxoglutarate ferredoxin reductase that are important intermediates for amino acid degradation (e.g., glutamate). Only Atabeyarchaeia can convert fumarate to malate via fumarate hydratase. The only portion of TCA cycle transcribed in any genome is 2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase, which can produce reducing power in the form of NADH.

A clue suggesting that amino acids are an important resource for Atabeyarchaeia and Freyarchaeia is the high expression of genes for protein and peptide breakdown (Figure 3.2). All three organisms are predicted to have the capacity to break down fatty acids via beta oxidation including crotonate (short-chain fatty acid) via the poorly described crotonate pathway. Furthermore they encode some enzymes involved in fermenting amino acids to H+, ammonium, acetate, and NAD(P)H via the hydroxyglutarate pathway (Table S7). The genomes also encode amino acid transporters and these are also highly transcribed in both archaeal groups. The ability to anaerobically degrade amino acids is consistent with predictions of the metabolism of the last Asgard common ancestor (Eme et al. 2023; Imachi et al. 2020).

Additionally, Freyarchaeia and Atabeyarchaeia can reverse the step in the formyl branch of the WLP that transforms glycine into methylenetetrahydrofolate (methylene-THF). Methylene-THF may then be converted to methyl-THF and then to formyl-THF, producing reducing power **(**Figure 3.2). Ultimately, the methyl group may be used to form acetate via the WLP. Interestingly Atabeyarchaeia-2 and Freyarchaeia expressed methylenetetrahydrofolate reductase (MTHFR) that is homologous to the enzyme used in the bacterial WLP and also plays a role in folate biosynthesis.

Environmental protection and adaptations

We predict that Atabeyarchaeia and Freyarchaeia are anaerobes expressing genes that encode oxygen-sensitive enzymes and proteins that protect against oxidative and other environmental stressors. Interestingly, all three organisms encode an ancestral version of clade I catalases (KatE) (Figure S18), Fe-Mn superoxide dismutase (SOD2), and unique to Freyarchaeia, a catalase-peroxidase (Figure S19) for protection against reactive oxygen species (ROS) (Figure S20). Previous analyses have described these expressed enzymes in acetogenic and sulfate-reducing bacteria and methanogenic archaea, but to our knowledge, not in Asgard archaea, indicating a potential adaptation to soil environments (Brioukhanov and Netrusov 2004). We also identified transcription for other environmental and stress responses, including transporters (e.g., nickel, arsenite, magnesium, iron, and copper), and heat shock proteins.

We infer that Atabeyarchaeia and Freyarchaeia use selenocysteine (Sec), the 21st amino acid, due to the presence of the Sec-specific elongation factor and Sec tRNA in their genomes. Additional Sec components, including phosphoseryl-tRNA kinase (Pstk), Sec synthase (SecS), selenophosphate synthetase (SPS) genes, and multiple Eukaryotic-like Sec Insertion Sequences are also present (Table S10). Phylogenetic analysis shows that the Sec elongation factor sequences from Atabeyarchaeia and Freyarchaeia are closely related to other Asgard members and Eukaryotes (Figure S21). We identified multiple selenoproteins encoded within each genome, including CoB-CoM heterodisulfide reductase iron-sulfur subunit (HdrA), peroxiredoxin family protein (Prx-like), selenophosphate synthetase (SPS), and the small subunit (~50 aa) of NiFeSec (VhuU). In VhuU, Sec plays a crucial role in mitigating oxidative stress. Sec can also enhance the catalytic efficiency of redox proteins, and the identified selenoproteins have the characteristic CXXU or UXXC sequence (Table S11) observed in redox-active motifs.

3.4 Discussion

Here, we reconstructed and validated three complete Asgard archaeal genomes from wetland soils in which these archaea comprise less than 1% of complex microbial communities. We used these genomes to define their chromosome lengths, structure and replication modes. It is relatively common for authors to report circularized genomes as complete, but this may be erroneous due to the prominence of local assembly errors, chimeras, scaffolding gaps and other issues in de novo metagenome assemblies (Chen et al. 2020; Watson and Warr 2019). Our genomes were thoroughly inspected, corrected and vetted after circularization, steps previously described to complete genomes from metagenomes (Tyson et al. 2004). These complete genomes are one of the first manual curations of short-read metagenomic data verified entirely with long-read analysis (Oxford Nanopore and/or PacBio) and the first complete short-read environmental Asgard genomes. Two of these genomes are from Atabeyarchaeia, a previously undescribed Asgard group and the first complete genome for Freyarchaeia. We predict bidirectional replication in Freyarchaeia and Atabeyarchaeia, suggesting bidirectional replication could have been present in the last common ancestor of eukaryotes and archaea, potentially playing a role in the emergence of the complex cellular organization characteristic of eukaryotes.

Overall, prior studies predict that Asgard archaea degrade proteins, carbohydrates, fatty acids, amino acids, and hydrocarbons (MacLeod et al. 2019; Seitz et al. 2016; Liu et al. 2018; Zhang et al. 2021). Lokiarchaeales, Thorarchaeia, Odinarchaeia, and Heimdallarchaeia are primarily organoheterotrophs with varying capacities to consume and produce hydrogen (Spang et al., 2019). Helarchaeales are proposed to anaerobically oxidize hydrocarbons (Seitz et al. 2019; Zhao and Biddle 2021; Zhang et al. 2021), whereas Freyarchaeia and Sifarchaeia are predicted to be heterorganotrophic acetogens capable of utilizing methylated amines (Sun et al. 2021; Farag Ibrahim F. et al. 2021). Hermodarchaeia are proposed to degrade alkanes and aromatic compounds via the alkyl/benzyl-succinate synthase and benzoyl-CoA pathway (Zhang et al. 2021). Gerdarchaeales may be facultative anaerobes and utilize both organic and inorganic carbon (Cai et al. 2020). Atabeyarchaeia and Freyarchaeia share several metabolic pathways with new lineages from the Asgard sister-clade TACK (e.g., Brock- and Culexarchaeia) and other deeply branching Asgard lineages. Based on the genomic and metatranscriptomic analyses, we predict that the soil-associated Atabeyarchaeia and Freyarchaeia are chemoheterotrophs that likely degrade amino acids and other carbon compounds. Both encode the EMP Pathway for cellular respiration and the WLP for CO_2 fixation.

Although Atabeyarchaeia and Freyarchaeia share key central metabolic pathways, they differ in that Freyarchaeia can metabolize compounds such as formaldehyde (C1), glycerol (C3), ribulose (C5), and glucose (C6), whereas Atabeyarchaeia can only metabolize C1, C3 and C5 compounds (Figure 3.2). The ability to metabolize C3 and C5 compounds is rare in Asgard archaea. While the entry points into the EMP pathway differ between the two, both exhibit the genetic repertoire necessary for converting carbohydrates into acetate. Both Atabeyarchaeia and Freyarchaeia may also be capable of growth as anaerobic acetogens via acetate production through the WLP. Similar to other Asgard archaea, they have methyltransferase complexes involved in the catabolism of quaternary amines (or yet unknown methylated substrates). Through methylated compounds, they may compete with methanogens and other anaerobic methylotrophic groups that rely on these substrates for methane production. These results align with recent studies suggesting a broader presence of methylotrophic metabolisms among archaea (De Anda et al. 2021; Kohtz et al. 2022; Zhang et al. 2021). It also opens up avenues for exploring the environmental impact of these metabolisms, particularly in relation to carbon cycling and greenhouse gas emissions (Offre et al. 2013).

Of particular interest is the predicted metabolic capability of Atabeyarchaeia and Freyarchaeia to degrade aldehydes. Aldehydes in soils come from several sources, including the microbial breakdown of methanol potentially produced from methane oxidation, degradation of plant and animal compounds, and products of industrial combustion and wildfires (e.g., volatile organic compounds). In fact, the California wetland soil that hosts these archaea contain charcoal, likely produced by wildfires. They are also predicted to be capable of growing on glycerol under anaerobic conditions capacity previously undescribed in Asgard archaea. Glycerol may be present in soil by the lysis of bacteria, yeast, and methanogenic archaeal cells that use glycerol as a solute, or by microbial fermentation of plant and animal triglycerides and phospholipids (Unden & Bongaerts, 1997). The presence of glycerol kinase and the respiratory glycerol-3-phosphate dehydrogenase (GlpABC) in Atabeyarchaeia and Freyarchaeia indicates these archaea might use with glycerol or glycerol-3-phosphate and fumarate as the terminal electron acceptor associated with proton translocation. This finding suggests a broader role for glycerol in Asgard archaeal energy metabolism and points to a possible conservation of this mechanism across different anaerobic environments. Understanding how these archaea metabolize glycerol will enhance our knowledge of their ecological roles and contributions to the carbon cycle in wetland ecosystems. Atabeyarchaeia and Freyarchaeia also produce and consume small organic molecules and H_2 that serve as substrates for methane production by methanogens that coexist in wetland soil.

The soil Asgard archaea encodes group 3c [NiFe]-hydrogenase genes, which were shown to be highly expressed *in situ*. Under specific conditions, autotrophic growth is likely supported by H_2 oxidation via the WLP. The presence of syntenic blocks encoding heterodisulfide reductase complexes, [NiFe] hydrogenases, and ATP synthase suggests a sophisticated apparatus for energy transduction, resembling mechanisms previously characterized in other archaeal groups (Sousa et al., 2016). Additionally, our results suggest the existence of an electron bifurcation mechanism in both Asgard archaea lineages, where electrons can be transferred from H_2 to ferredoxin and an unidentified heterodisulfide intermediate (Spang et al., 2019). Atabeyarchaeia and Freyarchaeia also have membrane-bound group 4 [NiFe]-hydrogenases that likely facilitate the oxidation of reduced ferredoxin generated through fermentative metabolism. However, this complex is novel in that it includes a HL helix on the L-like subunit and two antiporters, neither of which are part of biochemically characterized group 4 respiratory hydrogenases. The functional modeling of these complexes reveals structural congruences with known respiratory enzymes, hinting at a potential for chemiosmotic energy conservation that may be a widespread feature among the Asgard clade. The findings indicate a potential evolutionary connection between hydrogenases and complex I, aligning with the hypothesis that complex I may have evolved from ancestral hydrogenases (Friedrich and Scheide 2000; Efremov and Sazanov 2011).

These complete genomes provide insight into the unique metabolic pathways of Asgard archaea in soil environments, previously missed in primarily sediment-based descriptions. Of particular interest is the identification of genes encoding enzymes for oxidative stress response in both Atabeyarchaeia and Freyarchaeia, despite their anaerobic nature. The use of selenocysteine in key enzymes may provide another mechanism for dealing with increased oxidative stress. These Asgardarchaeota genomes suggest an adaptation to transient oxidative conditions in soil environments and additional competition for methanogenesis and anaerobic methyltrophy substrates.

3.5 Conclusion

We manually curated three complete genomes for Asgard archaea from wetland soils, uncovering bidirectional replication and an unexpected abundance of introns in tRNA genes. These features suggest another facet of the evolutionary relationship between archaea and eukaryotes. Metabolic reconstruction and metatranscriptomic measurements of *in situ* activity revealed a non-methanogenic, acetogenic lifestyle and a diverse array of proteins likely involved in energy conservation. The findings point to metabolic flexibility and adaptation to the dynamic soil conditions of wetlands. Finally, they contribute to cycling of carbon compounds that are relevant for methane production by coexisting methanogenic archaea.

3.6 Figures

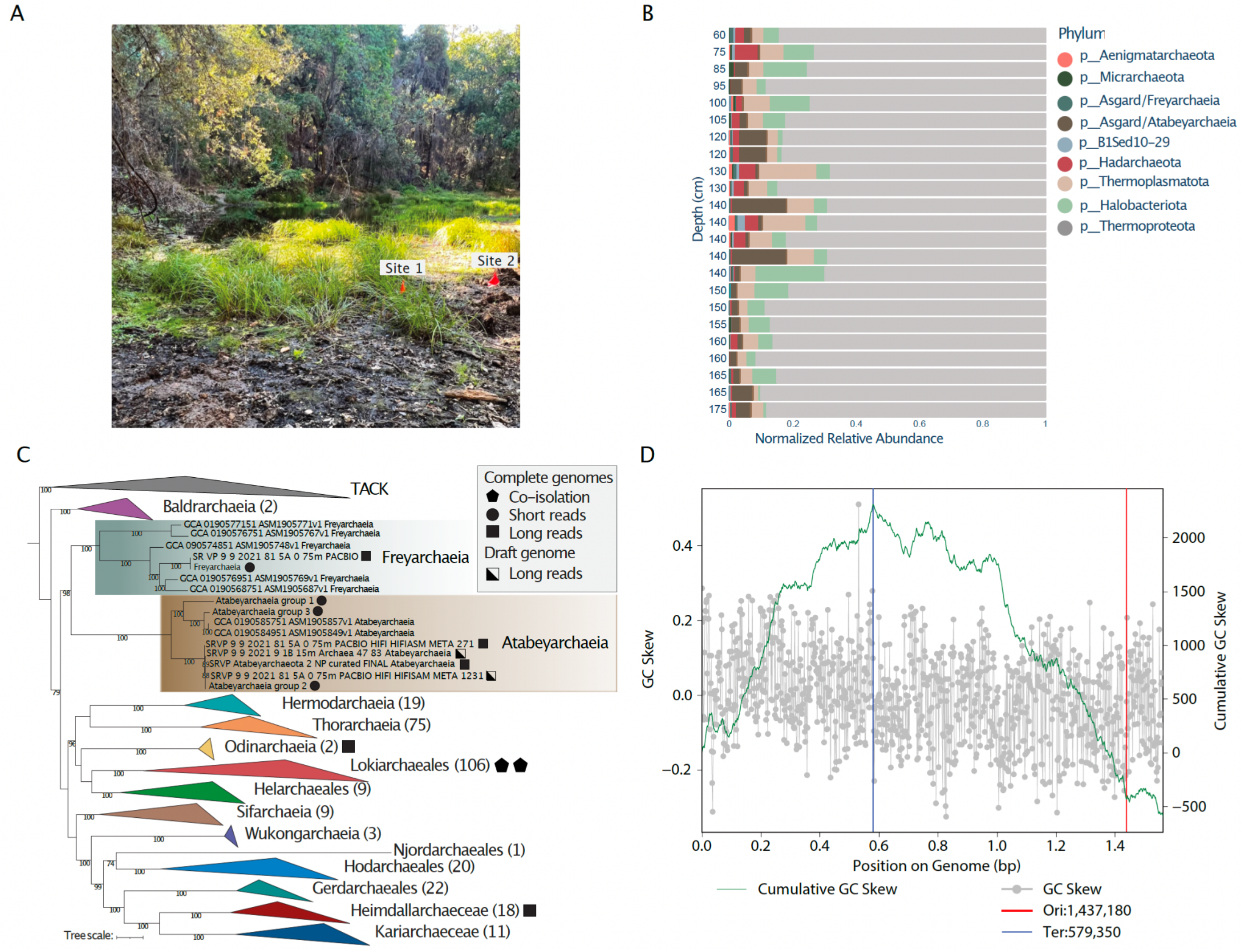


Figure 3.1 Archaea dominate deep regions of wetland soil and host novel Asgard archaea. A) Photograph of the vernal pool that was metagenomically sampled in this study, in Lake County, California, USA. B) Archaeal genomic abundance excluding bacterial genomes. C) Phylogenetic distribution of Asgard Archaea complete genomes. The maximum-likelihood phylogeny was generated with Iqtree v1.6.1, utilizing 47 concatenated archaeal Clusters of Orthologous Groups of proteins (arCOGs). The best-fit model was determined as LG+F+R10 based on the Bayesian Information Criterion. Non-parametric bootstrapping was conducted with 1,000 replicates for robustness. The filled-in square, circle, and triangle indicate closed complete genomes from short reads, published complete genomes from long reads, and genomes from co-isolated cultured representatives, respectively. The pentagon highlights the long read draft genomes from this site (PacBio or Nanopore). D) Bidirectional replication indication in Atabeyarchaeia complete genomes. The GC skew is shown as a grey plot overlaying the cumulative GC skew, presented as a green line. The blue lines mark the predicted replication terminus.

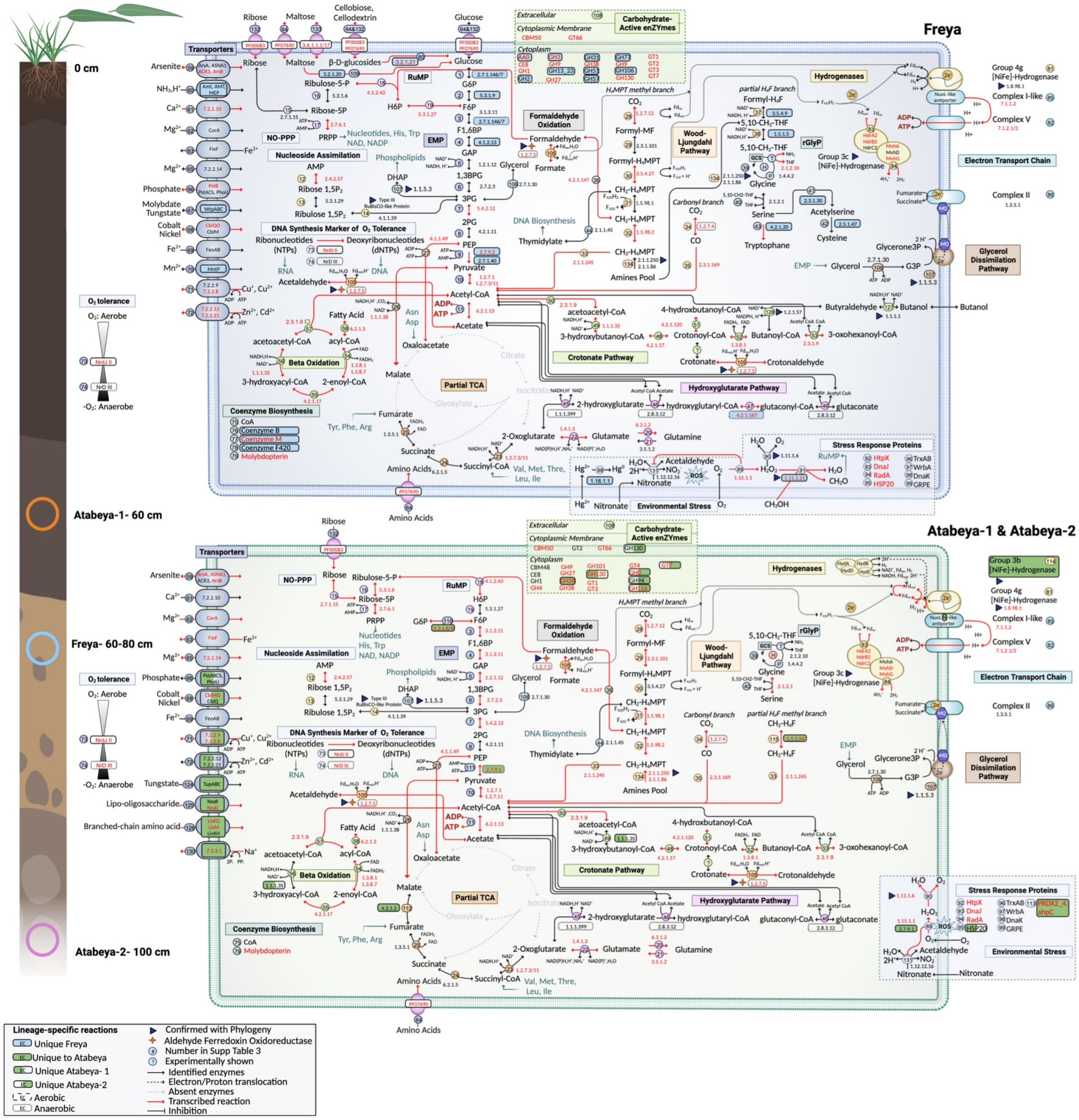


Figure 3.2 Metabolic capacities of terrestrial Atabeyarchaeia and Freyarchaeia for overall implications for biogeochemical cycling in wetlands. Inference of the pathways from the complete genomes is based on the comparison of predicted proteins with a variety of functional databases ("see materials and methods"). The extraction depth location within the cores is shown on the left. All reactions are numbers and correspond to Table S7. EC/TCDB numbers shaded fully or partially in blue or green are unique to the lineages and complete genomes, whereas the dashed boxes distinguish oxygen-sensitive enzymes. The multi-functional aldehyde ferredoxin oxidoreductase is shown with a star. Proteins marked with a triangle have generated phylogenies to determine their evolutionary histories and substrate specificity. Reactions with mapped transcripts are denoted with red text and arrows. Created using BioRender.com.

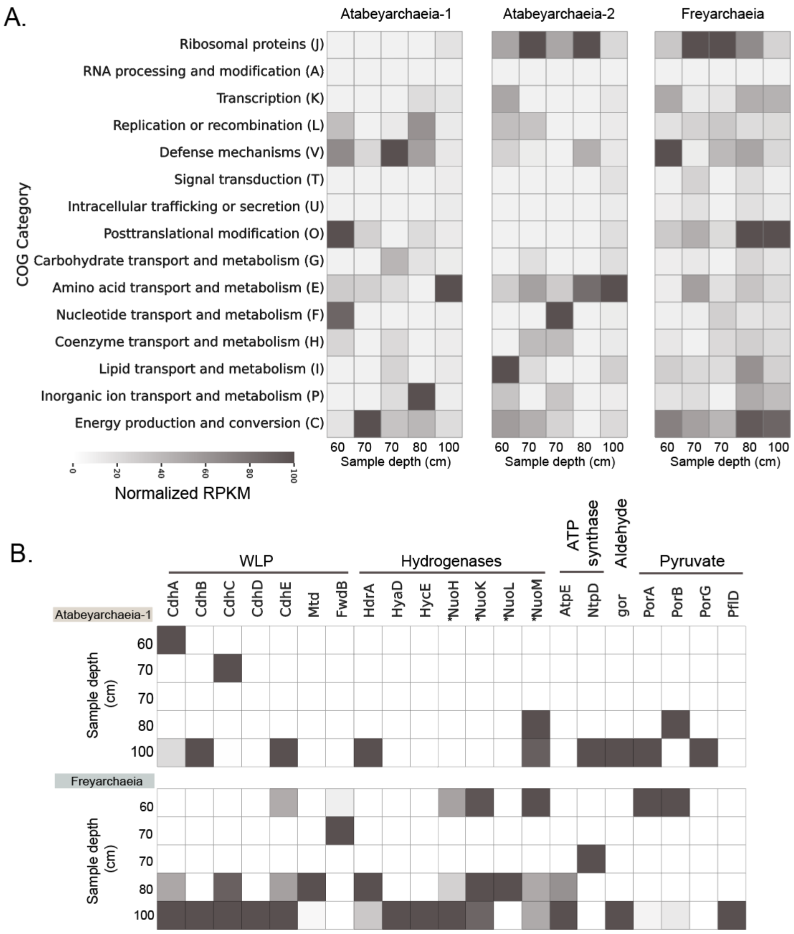


Figure 3.3 Metatranscriptomic profiling of soil-associated Asgard archaeal genomes. A) Heatmap visualization of normalized Reads Per Kilobase per Million mapped reads (RPKM) values for ORFs with high sequence similarity (≥95%) to the genomes of Atabeyarchaeia-1, Atabeyarchaeia-2, and Freyarchaeia, across various soil depths. A total of 2,191 open reading frames (ORFs) were categorized using the Clusters of Orthologous Groups (COG) database, with Atabeya-1, Atabeya-2, and Freya expressing 465, 804, and 922 unique ORFs, respectively. The ORFs were annotated and assigned to 15 COG categories, indicating the functional potential of each archaeal genome *in situ*. Columns represent metatranscriptomes from different soil depths, highlighting the spatial variability in the expression of key metabolic and cellular processes. B) Expanded heatmap of Atabeyarchaeia-1 and Freyarchaeia expressed genes under the category C: Energy production and conversion. Key genes of the WLP (CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase; *fwdB*, formate dehydrogenase; mtd, 5,10-methylene-H4-methanopterin dehydrogenase), hydrogenases and associated genes (HdrA, heterodisulfide reductase and group NiFe-hydrogenase; Mvh, methyl viologen reducing hydrogenase); HyaD (NiFe-hydrogenase_maturation_factor); HycE and Nuo like subunits, (group 4 NiFe-hydrogenase), ATP synthase (AtpE, V/A-type H+/Na+-transporting ATPase subunit_K; NtpD, V/A-type H+/Na+ transporting ATPase subunit D) and aldehyde metabolism (gor, Aldehyde:ferredoxin oxidoreductases), pyruvate oxidation (porABCD, 2-pyruvate:ferredoxin oxidoreductase; pflD, pyruvate-formate lyase).

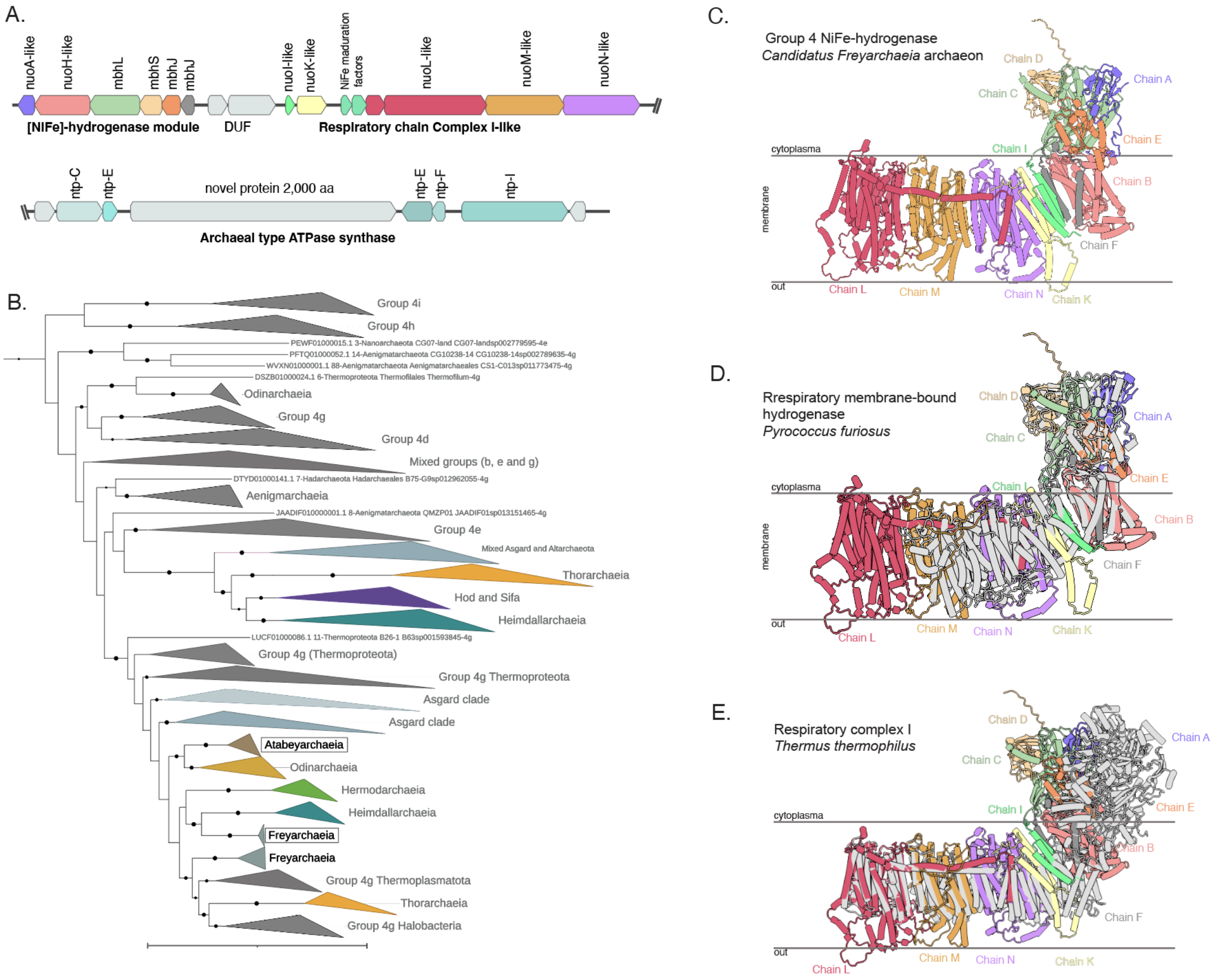


Figure 3.4 Phylogeny, genetic organization and structure of the novel group 4 energy-conservation complex I-like NiFe-hydrogenase from Asgard archaea. A) Genetic organization of the group 4 [NiFe]-hydrogenase module, the proton-translocating membrane module, and ATP synthase from the Freyarchaeia genome. B) Maximum likelihood phylogeny of group 4 [NiFe]-hydrogenase large subunit from Asgard archaea and reference sequences. The bolded taxonomic groups highlight the clades with genomes from this study used for modeling. C) AlphaFold models of [NiFe]-hydrogenase module and the proton-translocating membrane module where each candidate subunit is represented by a different color based on the best subunit matched. D) AlphaFold model of Freyarchaeia hydrogenase complex colored by chains, aligned with cryoEM structure of a respiratory membrane-bound hydrogenase (MBH) from Pyrococcus furiosus(Yu et al., 2018) (PDB ID: 5L8X). E) AlphaFold model of Freyarchaeia hydrogenase complex colored by chains, aligned with Crystal structure of respiratory complex I from *Thermus thermophilus* (Baradaran et al., 2013) (PDB: 4HEA).

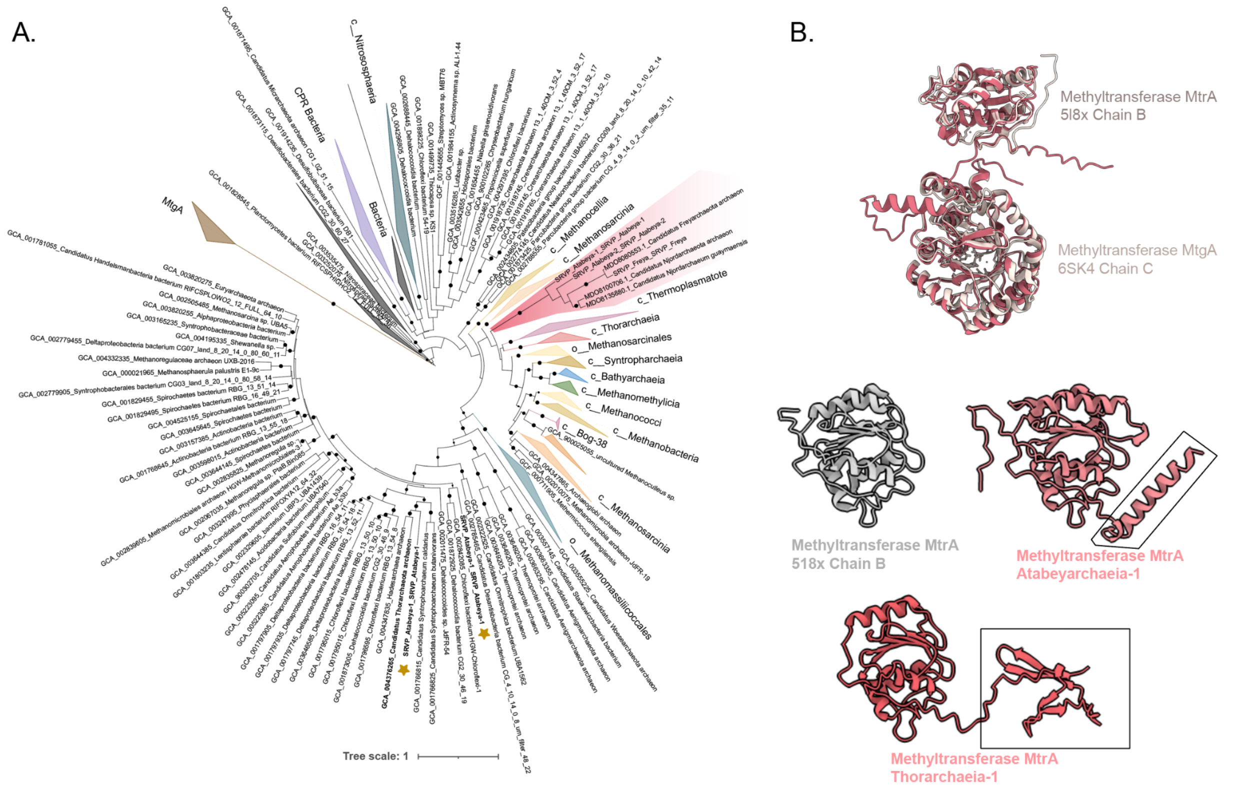


Figure 3.5 Non-methanogenic MtrA, MtrH and MtrAH fusion methyltransferases. A) Maximum likelihood phylogeny of MtrA and the MtrAH fusion, with reference to Tetrahydromethanopterin S-methyltransferase subunit A (MtrA) with the closest corresponding domains being MtrA from the characterized Tetrahydromethanopterin S-methyltransferase subunit A (MtrA) protein (PDB ID: 5L8X) (Adam et al., 2022). The coral colored clade is the novel fusion present in Atabeyarchaeia, Freyarcheia and other Asgardarchaeota members. B) AlphaFold models of Atabeyarchaeia-1 MtrAH (fusion) in coral aligned with the grey corresponding domains of the characterized protein Tetrahydromethanopterin S-methyltransferase subunit A (MtrA) (PDB ID: 5L8X)(Wagner et al., 2016) and Methyltransferase (MtgA) from *Desulfitobacterium hafniense* in complex with methyl-tetrahydrofolate (PDB ID: 6SK4) at the N terminus. We also modeled the putative MtrA present in Atabeyarchaeia-1 with the closest corresponding domains being MtrA from the characterized Tetrahydromethanopterin S-methyltransferase subunit A (MtrA) protein (PDB ID: 5L8X).

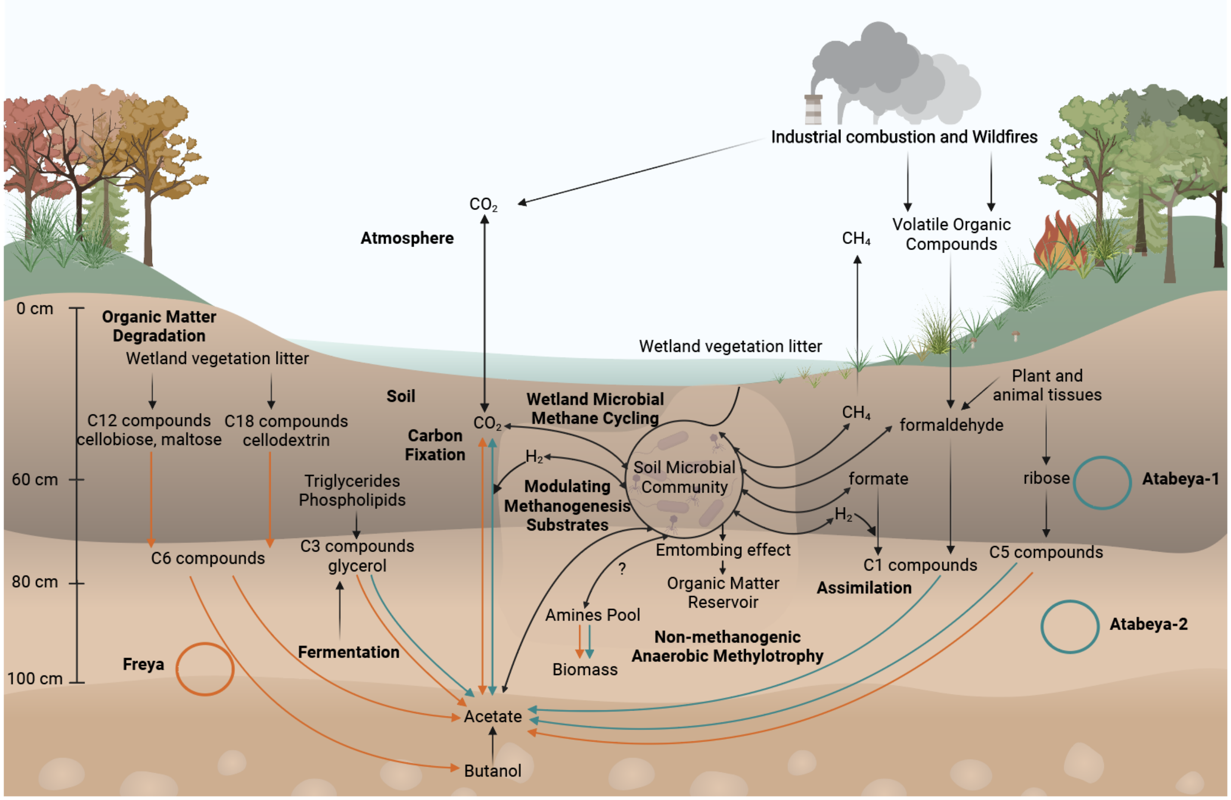


Figure 3.6 Overview of the wetland soil dynamics and biogeochemical cycling in Atabeyarchaeia and Freyarchaeia. Complete genomes for Atabeyarchaeia and Freyarchaeia are shown with green and orange circles, respectively. 2 Atabeyarchaeia genomes (Atabeya-1 and Atabeya-2) and 1 Freyarchaeia (Freya) genome were isolated and carefully curated and closed from wetland soil between 60-100 cm. These anaerobic lineages were shown in this study to encode the Wood-Ljungdahl Pathway for CO_2 fixation (e.g. methylated compounds such as quaternary amines) and EMP Pathway, components of chemolithotrophy and heterotrophy, producing acetate shown in arrows (green and orange), corresponding to the genome colors. Additionally, these lineages are involved in modulating methanogenesis substrates in these wetland soils. Detailed description of the specific pathways is found in main text, Figure 3.2, and supplementary materials. Created using BioRender.com.

3.7 References

Adam, P. S., Kolyfetis, G. E., Bornemann, T. L. V., Vorgias, C. E., & Probst, A. J. (2022). Genomic remnants of ancestral methanogenesis and hydrogenotrophy in Archaea drive anaerobic carbon cycling. *Science Advances*, *8*(44), eabm9651.

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., … Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, *578*(7795), 425–431.

Al-Shayeb, B., Schoelmerich, M. C., West-Roberts, J., Valentin-Alvarado, L. E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M. J., Williams, K. H., Doudna, J. A., & Banfield, J. F. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature*, *610*(7933), 731–736.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* , *36*(7), 2251–2252.

Arndt, F., Schmitt, G., Winiarska, A., Saft, M., Seubert, A., Kahnt, J., & Heider, J. (2019). Characterization of an Aldehyde Oxidoreductase From the Mesophilic Bacterium Aromatoleum aromaticum EbN1, a Member of a New Subfamily of Tungsten-Containing Enzymes. *Frontiers in Microbiology*, *10*, 71.

Baker, B. J., Saw, J. H., Lind, A. E., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., & Ettema, T. J. G. (2016). Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nature Microbiology*, *1*, 16002.

Baradaran, R., Berrisford, J. M., Minhas, G. S., & Sazanov, L. A. (2013). Crystal structure of the entire respiratory complex I. *Nature*, *494*(7438), 443–448.

Beam, J. P., Jay, Z. J., Schmid, M. C., Rusch, D. B., Romine, M. F., Jennings, R. de M., Kozubal, M. A., Tringe, S. G., Wagner, M., & Inskeep, W. P. (2016). Ecophysiology of an uncultivated lineage of Aigarchaeota from an oxic, hot spring filamentous "streamer" community. *The ISME Journal*, *10*(1), 210–224.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bevers, L. E., Bol, E., Hagedoorn, P.-L., & Hagen, W. R. (2005). WOR5, a novel tungsten-containing aldehyde oxidoreductase from Pyrococcus furiosus with a broad substrate Specificity. *Journal of Bacteriology*, *187*(20), 7056–7061.

Bojanova, D. P., De Anda, V. Y., Haghnegahdar, M. A., Teske, A. P., Ash, J. L., Young, E. D., Baker, B. J., LaRowe, D. E., & Amend, J. P. (2023). Well-hidden methanogenesis in deep, organic-rich sediments of Guaymas Basin. *The ISME Journal*, *17*(11), 1828–1838.

Bräsen, C., Esser, D., Rauch, B., & Siebers, B. (2014). Carbohydrate metabolism in Archaea: current insights into unusual enzymes and pathways and their regulation. *Microbiology and Molecular Biology Reviews: MMBR*, *78*(1), 89–175.

Brioukhanov, A. L., & Netrusov, A. I. (2004). Catalase and superoxide dismutase: distribution, properties, and physiological role in cells of strict anaerobes. *Biochemistry. Biokhimiia*, *69*(9), 949–962.

Buckel, W., & Barker, H. A. (1974). Two pathways of glutamate fermentation by anaerobic bacteria. *Journal of Bacteriology*, *117*(3), 1248–1260.

Buckel, W., & Thauer, R. K. (2013). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na(+) translocating ferredoxin oxidation. *Biochimica et Biophysica Acta*, *1827*(2), 94–113.

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/servlets/purl/1241166

Cai, M., Liu, Y., Yin, X., Zhou, Z., Friedrich, M. W., Richter-Heitmann, T., Nimzyk, R., Kulkarni, A., Wang, X., Li, W., Pan, J., Yang, Y., Gu, J.-D., & Li, M. (2020). Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Science China. Life Sciences*, *63*(6), 886–897.

Cai, M., Richter-Heitmann, T., Yin, X., Huang, W.-C., Yang, Y., Zhang, C., Duan, C., Pan, J., Liu, Y., Liu, Y., Friedrich, M. W., & Li, M. (2021). Ecological features and global distribution of Asgard archaea. *The Science of the Total Environment*, *758*, 143581.

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096.

Chastain, C. J., Failing, C. J., Manandhar, L., Zimmerman, M. A., Lakner, M. M., & Nguyen, T. H. T. (2011). Functional evolution of C(4) pyruvate, orthophosphate dikinase. *Journal of Experimental Botany*, *62*(9), 3083–3091.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* , *36*(6), 1925–1927.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315–333.

Cole, S. T., Eiglmeier, K., Ahmed, S., Honore, N., Elmes, L., Anderson, W. F., & Weiner, J. H. (1988). Nucleotide sequence and gene-polypeptide relationships of the glpABC operon encoding the anaerobic sn-glycerol-3-phosphate dehydrogenase of Escherichia coli K-12. *Journal of Bacteriology*, *170*(6), 2448–2456.

Cordero, P. R. F., Bayly, K., Man Leung, P., Huang, C., Islam, Z. F., Schittenhelm, R. B., King, G. M., & Greening, C. (2019). Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *The ISME Journal*, *13*(11), 2868–2881.

Crits-Christoph, A. (n.d.). *filter_reads.py: Filters BAM files created by bowtie2 for better read mapping. Use for genomes from metagenomes*. Github. Retrieved November 17, 2023, from https://github.com/alexcritschristoph/filter_reads.py

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A., 4th, Bik, H. M., & Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*, e243.

De Anda, V., Chen, L.-X., Dombrowski, N., Hua, Z.-S., Jiang, H.-C., Banfield, J. F., Li, W.-J., & Baker, B. J. (2021). Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. *Nature Communications*, *12*(1), 2404.

Dobbek, H., Gremer, L., Kiefersauer, R., Huber, R., & Meyer, O. (2002). Catalysis at a dinuclear [CuSMo(O)OH] cluster in a CO dehydrogenase resolved at 1.1-Å resolution. *Proceedings of the National Academy of Sciences*, *99*(25), 15971–15976.

Dombrowski, N., Williams, T. A., Sun, J., Woodcroft, B. J., Lee, J.-H., Minh, B. Q., Rinke, C., & Spang, A. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nature Communications*, *11*(1), 3939.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195.

Efremov, R. G., & Sazanov, L. A. (2011). Structure of the membrane domain of respiratory complex I. *Nature*, *476*(7361), 414–420.

Efremov, R. G., & Sazanov, L. A. (2012). The coupling mechanism of respiratory complex I - a structural and evolutionary perspective. *Biochimica et Biophysica Acta*, *1817*(10), 1785–1795.

Ellenbogen, J. B., Jiang, R., Kountz, D. J., Zhang, L., & Krzycki, J. A. (2021). The MttB superfamily member MtyB from the human gut symbiont Eubacterium limosum is a cobalamin-dependent γ-butyrobetaine methyltransferase. *The Journal of Biological Chemistry*, *297*(5), 101327.

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), 992–999.

Farag Ibrahim F., Zhao Rui, & Biddle Jennifer F. (2021). "Sifarchaeota," a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methylotrophy. *Applied and Environmental Microbiology*, *87*(9), e02584–20.

Feng, X., Cheng, H., Portik, D., & Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, *19*(6), 671–674.

Friedrich, T., & Scheide, D. (2000). The respiratory complex I of bacteria, archaea and eukarya and its module common with membrane-bound multisubunit hydrogenases. *FEBS Letters*, *479*(1-2), 1–5.

Gottschalk, G., & Thauer, R. K. (2001). The Na(+)-translocating methyltransferase complex from methanogenic archaea. *Biochimica et Biophysica Acta*, *1505*(1), 28–36.

Graham, D. E., & White, R. H. (2002). Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. *Natural Product Reports*, *19*(2), 133–147.

Greening, C., Biswas, A., Carere, C. R., Jackson, C. J., Taylor, M. C., Stott, M. B., Cook, G. M., & Morales, S. E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *The ISME Journal*, *10*(3), 761–777.

Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., & Chan, M. K. (2002). A new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*, *296*(5572), 1462–1466.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., … Takai, K. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*, *577*(7791), 519–525.

Kang, D. D., 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*, *7*, e7359.

Kato, N., Yurimoto, H., & Thauer, R. K. (2006). The physiological role of the ribulose monophosphate pathway in bacteria and archaea. *Bioscience, Biotechnology, and Biochemistry*, *70*(1), 10–21.

Kohtz, A. J., Jay, Z. J., Lynes, M. M., Krukenberg, V., & Hatzenpichler, R. (2022). Culexarchaeia, a novel archaeal class of anaerobic generalists inhabiting geothermal environments. *ISME Communications*, *2*(1), 1–13.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103–1110.

Kountz, D. J., Behrman, E. J., Zhang, L., & Krzycki, J. A. (2020). MtcB, a member of the MttB superfamily from the human gut acetogen Eubacterium limosum, is a cobalamin-dependent carnitine demethylase. *Journal of Biological*. https://www.jbc.org/article/S0021-9258(17)50060-5/abstract

Lang, K., Schuldes, J., Klingl, A., Poehlein, A., Daniel, R., & Brunea, A. (2015). New mode of energy metabolism in the seventh order of methanogens as revealed by comparative genome analysis of "Candidatus methanoplasma termitum." *Applied and Environmental Microbiology*, *81*(4), 1338–1352.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Lie, T. J., Costa, K. C., Lupa, B., Korpole, S., Whitman, W. B., & Leigh, J. A. (2012). Essential anaplerotic role for the energy-converting hydrogenase Eha in hydrogenotrophic methanogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(38), 15473–15478.

Li, J., Kang, P. T., Jiang, R., Lee, J. Y., Soares, J. A., Krzycki, J. A., & Chan, M. K. (2023). Insights into pyrrolysine function from structures of a trimethylamine methyltransferase and its corrinoid protein complex. *Communications Biology*, *6*(1), 54.

Liu, Y., Makarova, K. S., Huang, W.-C., Wolf, Y. I., Nikolskaya, A. N., Zhang, X., Cai, M., Zhang, C.-J., Xu, W., Luo, Z., Cheng, L., Koonin, E. V., & Li, M. (2021). Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature*, *593*(7860), 553–557.

Liu, Y., Zhou, Z., Pan, J., Baker, B. J., Gu, J.-D., & Li, M. (2018). Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *The ISME Journal*, *12*(4), 1021–1031.

Li, W., O'Neill, K. R., Haft, D. H., DiCuccio, M., Chetvernin, V., Badretdin, A., Coulouris, G., Chitsaz, F., Derbyshire, M. K., Durkin, A. S., Gonzales, N. R., Gwadz, M., Lanczycki, C. J., Song, J. S., Thanki, N., Wang, J., Yamashita, R. A., Yang, M., Zheng, C., … Thibaud-Nissen, F. (2021). RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*, *49*(D1), D1020–D1028.

MacLeod, F., Kindler, G. S., Wong, H. L., Chen, R., & Burns, B. P. (2019). Asgard archaea: Diversity, function, and evolutionary implications in a range of microbiomes. *AIMS Microbiology*, *5*(1), 48–61.

Marck, C., & Grosjean, H. (2003). Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA* , *9*(12), 1516–1531.

Mariotti, M., Lobanov, A. V., Guigo, R., & Gladyshev, V. N. (2013). SECISearch3 and Seblastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids*

Research, *41*(15), e149.

Mariotti, M., Lobanov, A. V., Manta, B., Santesmasses, D., Bofill, A., Guigó, R., Gabaldón, T., & Gladyshev, V. N. (2016). Lokiarchaeota Marks the Transition between the Archaeal and Eukaryotic Selenocysteine Encoding Systems. *Molecular Biology and Evolution*, *33*(9), 2441–2453.

Marreiros, B. C., Batista, A. P., Duarte, A. M. S., & Pereira, M. M. (2013). A missing link between complex I and group 4 membrane-bound [NiFe] hydrogenases. *Biochimica et Biophysica Acta*, *1827*(2), 198–209.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K., & Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, *41*(14), e140.

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science: A Publication of the Protein Society*, *32*(11), e4792.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682.

Mukund, S., & Adams, M. W. (1995). Glyceraldehyde-3-phosphate ferredoxin oxidoreductase, a novel tungsten-containing enzyme with a potential glycolytic role in the hyperthermophilic archaeon Pyrococcus furiosus. *The Journal of Biological Chemistry*, *270*(15), 8389–8392.

Müller, V. (2019). New Horizons in Acetogenic Conversion of One-Carbon Substrates and Biological Hydrogen Storage. *Trends in Biotechnology*, *37*(12), 1344–1354.

Nakamaru-Ogiso, E., Kao, M.-C., Chen, H., Sinha, S. C., Yagi, T., & Ohnishi, T. (2010). The membrane subunit NuoL(ND5) is involved in the indirect proton pumping mechanism of Escherichia coli complex I. *The Journal of Biological Chemistry*, *285*(50), 39070–39078.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), 555–560.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834.

Offre, P., Spang, A., & Schleper, C. (2013). Archaea in biogeochemical cycles. *Annual Review of Microbiology*, *67*, 437–457.

Okamura-Ikeda, K., Ohmura, Y., Fujiwara, K., & Motokawa, Y. (1993). Cloning and nucleotide sequence of the gcv operon encoding the Escherichia coli glycine-cleavage system. *European Journal of Biochemistry / FEBS*, *216*(2), 539–548.

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, *11*(12), 2864–2868.

Orsi, W. D., Vuillemin, A., Rodriguez, P., Coskun, Ö. K., Gomez-Saez, G. V., Lavik, G., Mohrholz, V., & Ferdelman, T. G. (2020). Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nature Microbiology*, *5*(2), 248–255.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., … Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427.

Pedroni, P., Della Volpe, A., Galli, G., Mura, G. M., Pratesi, C., & Grandi, G. (1995). Characterization of the locus encoding the [Ni-Fe] sulfhydrogenase from the archaeon Pyrococcus furiosus: evidence for a relationship to bacterial sulfite reductases. *Microbiology*, *141 (Pt 2)*, 449–458.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* , *28*(11), 1420–1428.

Picking, J. W., Behrman, E. J., Zhang, L., & Krzycki, J. A. (2019). MtpB, a member of the MttB superfamily from the human intestinal acetogen Eubacterium limosum, catalyzes proline betaine demethylation. *The Journal of Biological Chemistry*, *294*(37), 13697–13707.

Rodrigues-Oliveira, T., Wollweber, F., Ponce-Toledo, R. I., Xu, J., Rittmann, S. K.-M. R., Klingl, A., Pilhofer, M., & Schleper, C. (2023). Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature*, *613*(7943), 332–339.

Santesmasses, D., Mariotti, M., & Guigó, R. (2017). Computational identification of the selenocysteine tRNA (tRNASec) in genomes. *PLoS Computational Biology*, *13*(2), e1005383.

Santesmasses, D., Mariotti, M., & Guigó, R. (2018). Selenoprofiles: A Computational Pipeline for Annotation of Selenoproteins. *Methods in Molecular Biology* , *1661*, 17–28.

Schoelmerich, M. C., & Müller, V. (2019). Energy conservation by a hydrogenase-dependent chemiosmotic mechanism in an ancient metabolic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(13), 6329–6334.

Scott, I. M., Rubinstein, G. M., Lipscomb, G. L., Basen, M., Schut, G. J., Rhaesa, A. M., Lancaster, W. A., Poole, F. L., 2nd, Kelly, R. M., & Adams, M. W. W. (2015). A New Class of Tungsten-Containing Oxidoreductase in Caldicellulosiruptor, a Genus of Plant Biomass-Degrading Thermophilic Bacteria. *Applied and Environmental Microbiology*, *81*(20), 7339–7347.

Seitz, K. W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J. R., Teske, A. P., Ettema, T. J. G., & Baker, B. J. (2019). Asgard archaea capable of anaerobic hydrocarbon cycling. *Nature Communications*, *10*(1), 1822.

Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., & Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *The ISME Journal*, *10*(7), 1696–1705.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883–8900.

sickle: Windowed Adaptive Trimming for fastq files using quality. (n.d.). Github. Retrieved November 17, 2023, from https://github.com/najoshi/sickle

Søndergaard, D., Pedersen, C. N. S., & Greening, C. (2016). HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports*, *6*, 34212.

Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N., & Martin, W. F. (2016). Lokiarchaeon is hydrogen dependent. *Nature Microbiology*, *1*, 16034.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179.

Spang, A., Stairs, C. W., Dombrowski, N., Eme, L., Lombard, J., Caceres, E. F., Greening, C., Baker, B. J., & Ettema, T. J. G. (2019). Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nature Microbiology*, *4*(7), 1138–1148.

Sugahara, J., Kikuta, K., Fujishima, K., Yachie, N., Tomita, M., & Kanai, A. (2008). Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Molecular Biology and Evolution*, *25*(12), 2709–2716.

Sun, J., Evans, P. N., Gagen, E. J., Woodcroft, B. J., Hedlund, B. P., Woyke, T., Hugenholtz, P., & Rinke, C. (2021). Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota lineages. *ISME Communications*, *1*(1), 30.

Sun, J., Evans, P. N., Gagen, E. J., Woodcroft, B. J., Hedlund, B. P., Woyke, T., Hugenholtz, P., & Rinke, C. (2022). Correction: Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota. *ISME Communications*, *2*(1), 1–1.

Sweeney, B. A., Hoksza, D., Nawrocki, E. P., Ribas, C. E., Madeira, F., Cannone, J. J., Gutell, R., Maddala, A., Meade, C. D., Williams, L. D., Petrov, A. S., Chan, P. P., Lowe, T. M., Finn, R. D., & Petrov, A. I. (2021). R2DT is a framework for predicting and visualising RNA secondary structure using templates. *Nature Communications*, *12*(1), 1–12.

Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E., & Scott, S. S. (2008). Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *Journal of Experimental Botany*, *59*(7), 1515–1524.

Tamarit, D., Caceres, E. F., Krupovic, M., Nijland, R., Eme, L., Robinson, N. P., & Ettema, T. J. G. (2022). A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nature Microbiology*, *7*(7), 948–952.

Ticak, T., Kountz, D. J., Girosky, K. E., Krzycki, J. A., & Ferguson, D. J., Jr. (2014). A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), E4668–E4676.

Tocchini-Valentini, G. D., Fruscoloni, P., & Tocchini-Valentini, G. P. (2011). Evolution of introns in the archaeal world. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(12), 4782–4787.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Unden, G., & Bongaerts, J. (1997). Alternative respiratory pathways of Escherichia coli: energetics and transcriptional regulation in response to electron acceptors. *Biochimica et Biophysica Acta*, *1320*(3), 217–234.

Wagner, T., Ermler, U., & Shima, S. (2016). MtrA of the sodium ion pumping methyltransferase binds cobalamin in a unique mode. *Scientific Reports*, *6*, 28226.

Wagner, T., Koch, J., Ermler, U., & Shima, S. (2017). Methanogenic heterodisulfide reductase (HdrABC-MvhAGD) uses two noncubane [4Fe-4S] clusters for reduction. *Science*, *357*(6352), 699–703.

Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction [Review of *Errors in long-read assemblies can critically affect protein prediction*]. *Nature Biotechnology*, *37*(2), 124–126.

White, R. H. (1985). Biosynthesis of coenzyme M (2-mercaptoethanesulfonic acid). *Biochemistry*, *24*(23), 6487–6493.

Williams, T. J., Allen, M., Tschitschko, B., & Cavicchioli, R. (2017). Glycerol metabolism of haloarchaea. *Environmental Microbiology*, *19*(3), 864–877.

Wu, F., Speth, D. R., Philosof, A., Crémière, A., Narayanan, A., Barco, R. A., Connon, S. A., Amend, J. P., Antoshechkin, I. A., & Orphan, V. J. (2022). Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularized Asgard archaea genomes. *Nature Microbiology*, *7*(2), 200–212.

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* , *32*(4), 605–607.

Xie, R., Wang, Y., Huang, D., Hou, J., Li, L., Hu, H., Zhao, X., & Wang, F. (2022). Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. *Science China. Life Sciences*, *65*(4), 818–829.

Yoshihisa, T. (2014). Handling tRNA introns, archaeal way and eukaryotic way. *Frontiers in Genetics*, *5*, 213.

Yu, H., Wu, C.-H., Schut, G. J., Haja, D. K., Zhao, G., Peters, J. W., Adams, M. W. W., & Li, H. (2018). Structure of an Ancient Respiratory System. *Cell*, *173*(7), 1636–1649.e16.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358.

Zhang, J.-W., Dong, H.-P., Hou, L.-J., Liu, Y., Ou, Y.-F., Zheng, Y.-L., Han, P., Liang, X., Yin, G.-Y., Wu, D.-M., Liu, M., & Li, M. (2021). Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *The ISME Journal*, *15*(6), 1826–1843.

Zhao, R., & Biddle, J. F. (2021). Helarchaeota and co-occurring sulfate-reducing bacteria in subseafloor sediments from the Costa Rica Margin. *ISME Communications*, *1*(1), 25.

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., & Anantharaman, K. (2022). METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, *10*(1), 33.

Transitional section

The complete genomes of soil-associated Asgard archaea like Atabeyarchaeia and Freyarchaeia, as described in Chapter 3, provide crucial insights into their metabolic capabilities and potential roles in terrestrial ecosystems. However, understanding the full evolutionary dynamics of these archaeal lineages, considered among the closest living relatives to the ancestors of eukaryotes, requires examining the mobile genetic elements (MGEs) that can facilitate horizontal gene transfer and genomic plasticity. In Chapter 4, I utilized these complete genomes and strain variants recovered using long-read sequencing to study the mobile genetic elements that associate with these Asgard archaea. Characterizing these MGEs illuminates the mechanisms by which Asgard archaea acquire new genes, defend against foreign genetic elements, and modulate gene expression through processes like methylation. Collectively, the genomes and associated MGEs shed light on the genetic versatility that may have contributed to the emergence of critical eukaryotic features during the process of eukaryogenesis from an Asgard archaeal ancestor.

4. Genetic elements and defense systems drive diversification and evolution in Asgard archaea

The following chapter is a modified version with permission of authors of the following published preprint: Valentin-Alvarado L, Ling-Dong S, Appler K, Lou YC, Crits-Christoph A, De Anda V, Leão L, Adler BA, Roberts RJ, Sachdeva R, Baker BJ, Savage DF, Banfield JF (2024) *In prep*. Genetic elements and defense systems drive diversification and evolution in Asgard archaea.

Abstract

Asgard Archaea are of great interest as the progenitors of Eukaryotes, but little is known about the mobile genetic elements (MGEs) that may shape their ongoing evolution. Here, we describe MGEs that replicate in Atabeyarchaeia, wetland Asgard archaea phylum represented by two complete genomes. We used soil depth-resolved population metagenomic datasets to track 18 MGEs for which genome structures were defined and precise chromosome integration sites could be identified for confident host linkage. Additionally, we identified a complete 20.67 kilobase pair (kbp) circular plasmid (the first reported for Asgard archaea) and two groups of viruses linked to Atabeyarchaeia, via CRISPR spacer targeting. Closely related 40 kbp viruses possess a hypervariable genomic region encoding combinations of specific genes for small cysteine-rich proteins structurally similar to restriction-homing endonucleases. One 10.9 kbp circularizable plasmid-like MGE integrates genomically into an Atabeyarchaeia chromosome and has a 2.5 kbp circularizable element integrated within it. The 10.9 kbp MGE encodes a highly expressed methylase with a sequence specificity matching an active methylation motif identified by PacBio sequencing. Restriction-modification of Atabeyarchaeia differs from that of another coexisting Asgard archaea Freyarchaeia which has few identified MGEs but possesses diverse defense mechanisms, including DISARM and Hachiman not found in Atabeyarchaeia. Overall, defense systems and methylation mechanisms of Asgard archaea likely modulate their interactions with MGEs, and integration/excision and copy number variation of MGEs in turn enable host genetic versatility.

All main figures for this manuscript can be found below in section 4.6. All supplementary files (including figures and tables) can be found online.

4.1 Introduction

Asgard archaea, including Loki-, Hermod-, Thor-, Odin-, Baldr-, Freya-, Sif-, Heimdall-, Atabey-, and Wukongarchaeia, bridge our understanding of the evolution of eukaryotes and prokaryotes. Their genomic features, particularly the presence of eukaryotic signature proteins (ESPs), provide insights into the steps leading to eukaryotic cellular complexity. Recent phylogenetic analyses place eukaryotes within Asgard archaea (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), most closely related to Hodarchaeales. Despite intense interest in their functionality and evolutionary relationships, little has been reported regarding Asgard mobile genetic elements (MGEs) that may shape their population diversity, contribute to genome divergence and facilitate cross-domain horizontal gene transfer[4]. Recent studies identified viruses of Loki-, Odin-, and Thor and Heimdallarchaeia (Medvedeva et al. 2022; Rambo et al. 2022; Tamarit et al. 2022; Wu et al. 2022), as well as putative transposons carrying cargo genes that replicate within Heimdallarchaeia (Wu et al. 2022), primarily based on CRISPR targeting. To our knowledge, no plasmids or plasmid-like elements have been described for Asgard archaea.

Recently, we reported two complete genomes for Atabeyarchaeia, a new group of Asgard archaea, and the first complete genome for Freyarchaeia (Valentin-Alvarado et al. 2023). Here, we track subtle strain variation of integrated MGEs over a soil depth profile by aligning the Illumina reads from a series of metagenomic samples from vernal pool wetland soil. We uncovered a suite of MGEs of Atabeyarchaeia that coexist in integrated and circular forms in metagenomic samples. Importantly, this approach establishes the host, and exactly defines MGE insertion sites and MGE lengths. Thus, we expand the repertoire of MGEs of Asgard archaea, explore their gene inventories, and shed light on MGE integration and excision events in natural poupulations of Atabeyarchaeia. Based on CRISPR targeting, we also genomically define groups of circular viruses and unclassified MGEs, some of which feature hypervariable regions enriched in cysteine-rich proteins predicted to be restriction endonucleases. We also describe genomically-encoded defense systems of both Atabeyarchaeia and Freyarchaeia that we confirm to be actively expressed with metatranscriptomic data. Leveraging the ability of PacBio SMRT sequencing data to identify DNA methylation sites, we report methylation patterns that distinguish these archaea, as well as a transcriptionally active MGE-encoded methylase that may enable the MGE to avoid restriction-based defense systems and compete with other MGEs. Overall, this study leverages complete metagenome-based genomes, read-based population analyses, metatranscriptomics and long read-based epigenetic analyses to provide a mosaic set of Asgard MGEs with likely relevance to Asgard archaeal evolution.

4.2 Materials and Methods

Sample acquisition, nucleic acid extraction and sequencing

DNA and RNA extractions and shotgun metagenome library construction were described in ref9,30. Briefly, we collected soil cores from a seasonally flooded wetland (SRVP) in Lake County, California, in October 2018, October 2019, November 2020 and October 2021 (38°41’39”N 122°31’36”W 571m). Samples were frozen in the field using dry ice, and kept at -80 C until extraction. The Qiagen PowerSoil Max DNA extraction kit was used to extract DNA from

5-10 g of soil, and the Qiagen AllPrep DNA/RNA extraction kit was used to extract RNA from 2 g of soil. Samples were sequenced by the QB3 sequencing facility at the University of California, Berkeley on a NovaSeq 6000. Read lengths for the 2018 DNA samples and the RNA samples were 2x150 bp and 2x250 bp for the 2019-2021 DNA samples. A sequencing depth of 10 Gb was targeted for each of 2018, 2020, and 2021 samples, and 20 Gbp for each of the 2019 samples. PacBio sequencing was obtained from a subset of deep soil samples from 2021 via the University of Maryland sequencing facility. Samples from September 9, 2021 from 140cm and 75cm were sequenced using a Sequel II to generate PacBio HiFi reads. Reads were quality trimmed using BBDuk (bbduk.sh minavgquality=20 qtrim=rl trimq=20) (Bushnell 2014)and assembled with hifiasm-meta (Feng et al. 2022).

Discovery of integrated genetic elements using closed complete genomes

Our manual approach to identifying integrated MGEs was based on 1) anomalously low or high coverage over a region, 2) short reads supporting the excision (see Figure 4.1), and 3) genomic architecture and gene annotations. To investigate the integration and excision state of the Yucahu plasmid, we employed a combination of sequencing and bioinformatic analyses. The metagenome reads were aligned to the reference genome sequences. Read coverage was employed to determine the relative abundance and distribution of each MGE across the different samples. Additionally, in-depth examination of the reads allowed for the identification of integration events, mini elements, and MGEs within the plasmid. To further characterize the MGE, we conducted comparative genomics analysis, comparing genomes lacking the MGE with those containing it. This enabled us to accurately determine the length of the MGE and pinpoint the specific sites of integration within the plasmid.

Identification and genome curation of Atabeyarchaeia-associated exogenous MGEs

We used metagenomic datasets to search for candidate mobile genetic elements associated with Atabeyarchaeia and Freyarchaeia. Screening was based on taxonomic profile, GC contents and CRISPR-based targeting (see below). All the candidate contigs were manually curated using Geneious Prime-2023.1.2. The manually curated genomes were de novo reconstructed from high-quality Illumina metagenomic data. Manual genome curation methods generally follow (Chen et al. 2020). Long-read PacBio data were used to verify and expand the sequence dataset. Replichores of complete genomes were predicted according to the GC skew and cumulative GC skew calculated by the iRep package (gc_skew.py). Complete MGE genomes with viral structural genes were classified as viruses and genomes that did not have viral structural genes were categorized as plasmid or other genetic elements such as transposons, conjugative elements, and unclassified.

CRISPR-Cas systems and classification of soil Asgard-associated viruses

CRISPR-Cas systems in Atabeyarchaeia and Freyarchaeia genomes were identified using CRISPRCasTyper v1.8.0 (Russel et al. 2020). Spacers were extracted from reads by mapping reads to the corresponding CRISPR arrays via BBMap (sourceforge.net/projects/bbmap/). Recruited spacers were matched against all assembled scaffolds with ≤1 mismatch using Bowtie v1.3.1 (Langmead et al. 2009). Scaffolds that are targeted by CRISPR spacers and not affiliated with microbial genomes were curated manually to completion. The phylogenetic classification was predicted based on genome-wide similarities using ViPTree whole proteome-based similarity of MGE from Atabeyarchaeia and other Asgardarchaeota (Table S9).

Coverage calculation of integrated genetic elements and host-chromosome

We aligned metagenome reads to reference genomes using the BBMap's short-read aligner. A minimum identity threshold of 0.95 was required, and ambiguously mapping reads were discarded. The coverage values of the MGEs, and total genome coverage values were calculated from the alignment/map (BAM) files. The positions of integrated elements were defined. For the genome coverage calculation, the entire genome was divided into two sections: the region before the start of the larger integrated element and the region after the end of the larger integrated element. This approach enabled systematic, efficient calculation of the coverage of integrated elements within genomes.

Methylation analysis via REBASE and Single Molecule, Real-Time (SMRT)

Methylation patterns within the genomes of Atabeyarchaeia and Freyarchaeia were investigated by mapping PacBio circular-consensus reads metagenomic reads to each of the three curated circular reference genomes for Atabeyarchaeia-1, Atabeyarchaeia-2, and Freyarchaeia using minimap2. The resulting BAM files were then processed and analyzed to identify methylation patterns using the ipdSummary and motifMaker commands in the SMRT Link analysis software package (v11.0; Pacific Biosciences, Menlo Park, CA, USA). To annotate methyltransferase (MTase) activities and restriction enzyme sites, the sequenced genomes and identified methylated motifs were compared against the Restriction Enzyme Database (REBASE) (Roberts et al. 2015; Blow et al. 2016). This comparison enabled the annotation of methylation sites and the determination of specific motifs associated with methyltransferase activity and restriction-modification systems within these archaeal genomes.

Phylogenetic analysis of hallmark proteins present in MGEs

We compiled the top 25 to 50 best matches for candidate proteins from the NCBI database, ggKbase and UniProt. The sequences were aligned using MAFFT with the parameters --localpair --maxiterate 1000 --reorder. Following alignment, each was trimmed using TrimAl, applying a gap threshold of 0.7. The final alignments underwent manual inspection using Geneious (see above). Maximum likelihood trees were inferred using IQ-TREE version 1.6.12, employing the auto option for model selection, a bootstrap value of 1000, and identifying the best-fit model for constructing the final trees. The details of all models used are included in the description of each figure. Trees were visualized using iTOL. All hallmark MGEs protein alignments and trees have been provided in the supplementary data for further reference.

Opia virus proteins structure predictions

Proteins were structurally modeled using AlphaFold2 (Jumper et al. 2021). Foldseek (van Kempen et al. 2023) was employed to identify structural homologs and structural alignments and comparisons to the modeled proteins were conducted using UCSF ChimeraX (Meng et al. 2023). Active site contents were derived from published descriptions for the reference structures.

Structural analyses and structural phylogeny of the ESP

The protein sequences of small GTPases found were analyzed along with eukaryotic small GTPases identified as sequence homologues (Figure S7-8). These sequences were submitted for structural modeling using ColabFold v1.5.240 (Mirdita et al. 2022). Multi-sequence alignments were performed using the MMseqs2 mode and the AlphaFold2_ptm models. Two recycling steps were employed to improve model prediction. The structural models were used as queries to search for structural homologues in the RCSB Protein Data Bank using FoldSeek easy-search feature, with a cutoff of >15% identity and <E-05 e-value.

Protein structures identified by FoldSeek were integrated with the models generated in ColabFold. A multi-structural alignment (MSTA) of these structures was carried out using the default parameters of mTM-align (Dong et al. 2018) (Figure S7-C). The resulting MSTA was further analyzed using IQ-TREE v2.0.3 (Model LG+I+G4 chosen according to BIC), yielding the dendrogram in Figure S7-B. The pairwise matrix obtained from the mTM-align process (Table S8) was utilized to select proteins suitable for 3D reconstruction of their alignments (Figure S7-D). This was done using the Needleman-Wunsch algorithm and the BLOSUM-62 matrix within the ChimeraX software (Meng et al. 2023)

4.3 Results

Complete Atabeyarchaeia genomes contain integrated genetic mobile elements

Short and long-read metagenomic and metatranscriptomic datasets were generated from wetland soil sampled at a single local site (Methods). We mapped reads from 28 samples collected from soil depths of 60 to 175 cm to the Atabeyarchaeia-1, Atabeyarchaeia-2 and Freyarchaeia genomes previously assembled from this environment and used mapped read details to uncover evidence for integrated, excised and coexisting circularized MGEs (Figure 4.1). Absence of MGEs in some cells can lead to lower coverage than average coverage over the integrated region, whereas higher read depth of coverage is likely due to coexisting extrachromosomal versions of the MGE (Kieft and Anantharaman 2022). By manual inspection of sequencing depth and read alignment discrepancies, we identified 14 chromosomally integrated genetic elements in the Atabeyarchaeia-1 (Atabeya-1) genome and 4 in the Atabeyarchaeia-2 (Atabeya-2) genome, ranging from 1.3 to 40 kbp in length. No integrated elements were identified in the Freyarchaeia genomes using this approach (Figure 4.2).

Of the 18 integrated MGEs in Atabeyarchaeia, five were classified as insertion sequence-like transposons (IS), three as putative integrative conjugative plasmids of 7.9 - 12 kbp in length carrying integration machinery and cargo genes, two as defense islands, six as pro-viruses, and two could not be classified (Table S1). To date, the only Asgard non-viral integrated genetic elements reported are Heimdallarchaeia “aloposons”8, which are transposons that carry cargo genes. These previously reported MGEs do not display any homology at the nucleotide level with those found in Atabeyarchaeia. However, some Atabeyarchaeia integrated genetic elements and these aloposons encode partition proteins (ParB-like) that are distantly related to tyrosine-like integrases (Figure S1-A). The Atabeyarchaeia tyrosine-like integrases are most closely related to those found in genomes of Njordarchaeales, Bathyarchaeia and Aenigmarchaeota, which share a similar ecological distribution in terrestrial wetlands and also occur in deep ocean sediments (Seitz et al. 2019) (Figure S1-B).

Ten of the integrated genetic elements coexist in circularized form with their integrated versions in the same metagenomic samples (e.g., Figure 4.1). One of these of particular interest is Atabeya-1 MGE-i (Yucahu-i, in homage to the son of the Taíno goddess Atabey—reflecting our previous designation of the host archaeon as Atabeyarchaeia)—for which the coexisting circularized version in 60 cm deep soil is four times more abundant than the integrated version (Figure 4.2A). Some of the reads span the genome indicate that a subset of Atabeya-1 genomes lack or have excised this integrated element (Figure 4.1; Figure 4.2B), enabling us to determine the exact length of the Yucahu-i to be 10,867 bp. The MGE is inserted following an AATTAACTTAT sequence that is also present at the end of the integrated Yucahu-i and occurs within the excised, circularized version. This region likely represents the attachment (att) site, a unique location within the genome. The low GC content (9%) compared to the genome-wide average (~50%), suggests that the DNA in this area may exhibit increased susceptibility to cleavage during processes such as excision or integration.

The Yucahu-i element includes 11 open reading frames (Figure 4.2A). Some of the gene products could be functionally annotated using protein homology and *in silico* structural prediction. The first gene encodes a tyrosine recombinase/integrase that likely recognizes and cuts at the AATTAACTTAT motif in the genome and in the circularized version (resulting in Yucahu-i linearization), and may be involved in integration of the linear sequence. The subsequent gene is a Holliday junction resolvase, which likely acts in conjunction with the integrase. We are uncertain if a host integration factor is required, but it is possible that two of the following genes predicted to encode DNA-binding proteins may have this function. Yucahu-i also encodes a superfamily 3 (SF3) helicase that may unwind the DNA and initiate plasmid replication11, and a Type II-G restriction-modification (IIG RM) protein fusion that combines endonuclease and methyltransferase activities. Phylogenetic analyses place the IIG RM sequence shares its most recent common ancestor with sequences found in DPANN archaeal genomes (>60% amino acid identity). Basal to this clade are many sequences from bacteria, which supports the inference that the origin of the sequences in question is likely archaeal, potentially acquired via horizontal gene transfer (Figure S2). Metatranscriptomic data indicate that the IIG RM gene is transcribed (Supplementary Data). Based on predicted protein functions and presence of the excised, circular (copy number up to 4x) mobile genetic element, Yucahu-i is likely a plasmid.

We investigated how frequently Yucahu-i was integrated in, or coexisted in circular form with, the Atabeya-1 genome by systematically analyzing reads from the 20 soil metagenomes that contained this archaeon (Figure S3, Table S2). In 11 samples, Yucahu-i is integrated into essentially all Atabeya-1 cells, however, the data indicate substantial variation in presence/absence of the integrated version and in the copy number of the circularized version (Figure 4.2, Figure S3, Table S2). A few reads from the 70 cm deep soil revealed evidence for the circularization of a 2,644 bp element that is integrated within Yucahu-i. We refer to this as mini-Yucahu-i (Figure 4.3C). Its presence highlights the genetic plasticity of the plasmid. The mini-Yucahu-i carries a putative ParG, a hypothetical protein, and Holliday junction ATP-dependent DNA helicase RuvB. Interestingly, the identical 11 bp Yucahu-i putative attachment motif is also present adjacent to, and within, a 7,848 bp integrated element in the Atabeya-2 genome (iMGE-xvi). However, the genomes share no detectable similarity, and the percentage of identity of the tyrosine integrases are < 25% and they are phylogenetically unrelated (Figure S4).

MGEs targeted by CRISPR systems

To explore exogenous MGEs of Atabeyarchaeia and Freyarchaeia, we mined CRISPR spacers from their genomes and matched them to unbinned metagenomic scaffolds from the same wetland soil. More than 30 putative MGE scaffolds are confidently targeted by CRISPR spacers and thus predicted to have once replicated within Atabeyarchaeia. We manually curated them and obtained one complete 20.8 kbp circular plasmid genome, two circular, complete 40 kbp genomes for a pair of closely related viruses, and a circular complete 26.7 kbp genome for an unclassified MGE.

The 20.8 kbp plasmid has 24 open reading frames (ORFs), primarily encoding hypothetical proteins (Figure S5, Table S3-S4). It also encodes plasmid proteins such as protein repressor ribbon-helix-helix protein from the copG family12, usually present in bacterial conjugative plasmids. Other predicted proteins are implicated in autonomous replication, such as a DNA primase-helicase and a tyrosine integrase, as well as other genes, involved in nucleic acid processing. Seven of these proteins contain transmembrane domains suggesting the presence of a putative conjugative system or a secretion-like system (Figure S5). A protein with a Glu-Glu motif was annotated as an integral membrane CAAX-like protease self-immunity, based on structural modeling and phylogeny (Figure S6). It encodes a CTPase with similar function to ParB, a protein typically associated with plasmid chromosome partitioning during replication. Phylogenetic analysis places this protein within a clade that contains MGEs recently discovered in Heimdallarchaeia (Figure S1-A). Interestingly, this clade also contains ParB-like proteins from draft genomes of Lokiarchaeia and Thorarchaeia, along with other archaeal genomes (e.g., Sulfulobales), suggesting that this plasmid lineage is widespread in other Asgard archaea. Also included are sequences from *Streptomyces* plasmids. Therefore, these plasmid partitioning genes may have undergone inter-domain horizontal gene transfer.

The circular 40,094 bp virus is predicted to infect Atabeyarchaeia-2 based on CRISPR spacer matches. We named this virus 'Opia,' after a mythical creature associated with the Taíno goddess Atabey. Interestingly, we found this virus integrated into the end of a 2.58 Mbp PacBio-derived genome fragment from an related Atabeya-2 strain (Atabeya-2'), confirming its host association. The genome encodes structural proteins such as a capsid-like protein, phage head morphogenesis protein, phage portal protein, tail-like structural proteins and a phage terminase large subunit (K06909:xtmB). Comparison of Opia terminase and capsid proteins with reference archaeal and bacterial virus proteins placed them with those of other Asgard viruses, including Nidhogg, a virus of Helarchaeales (Figure 4.4A-B). The Opia genome harbors genes for various nucleic acid processing proteins, such as a Mu-like prophage protein com and putative transposase, tyrosine recombinase-like, site-specific DNA-methyltransferase, and a ParB-like CTPase. A DNA polymerase sliding clamp subunit (PCNA) like protein (Figure 4.4C) potentially interacts with viral replication proteins, promoting viral DNA synthesis and possibly manipulating host cellular pathways to facilitate viral replication and immune evasion.

Phylogenetic analysis positioned the Opia PCNA-like sequence within an Asgard archaea clade with a homolog present in the chromosome of Atabeyarchaeia-2. The most similar homolog was a predicted protein from an Njordarchaeales (MCD6165036.1) from the Auka vent field13, suggesting that viruses related to Opia integrate into other Asgard genomes. Homologs of this protein occur in the Sköll viral genome that infects Lokiarchaeia5–7 as well as other archaeal viruses14,15. At the whole proteome level, Opia has similarity to the Ratatoskr, Nidhogg, Skoll, and Fenrir viruses6, known to infect various Asgard archaea (Figure 4.4D).

We identified at least seven distinct Opia virus variant genotypes. The sequences align near-perfectly over >85% of the genomes (Figure 4.5A). All Opia variants (and their identifiable fragments) are exactly targeted by one CRISPR spacer present in loci of both Atabeya-2 and Atabeya-2' (two identical sequential spacers in Atabeya-2'), despite the presence of the Opia provirus in the Atabeya-2' genome. The provirus was only partially recovered, so it is impossible to say whether the integrated version differs in the targeted region. A hotspot in the Opia virus genomes encodes a series of genes that are distinctly different in some variants. Included are up to six small cysteine-rich proteins, many with predicted double Zn-binding domains (Figure 4.5A). The genes for specific cysteine-rich proteins occur in different recombinations from different genotypes (e.g., one has sequence types A, B, D another A, C, D, and another, C, E). In addition, a three-gene block (one of which has sequence variants) and adjacent intergenic sequences are variably present/absent. Finally, different versions of ParB-like partition proteins occur in the variable region and some lack a C-terminal endonuclease domain (Figure 4.5).

We predicted the structures of the largest cysteine-rich protein from Opia-3708 (gene 46) and two Opia-19564 proteins (genes 16, 18). All three represent different protein sequence clusters but they share core structural components, including two sets of four cysteines, and an alpha helix in proximity to paired antiparallel beta strands (i.e., α,**ββ**,-metal; Figure 4.5C). HHpred predicts the Opia-3708 protein (197 aa) to be related to an HNH endonuclease and the best match for the three-dimensional structure (PDB 3M7K) is the Rare-Cutting HNH Restriction Endonuclease PacI, a homing endonuclease that is one of the smallest restriction endonucleases known (142 aa) (Shen et al. 2010). Zinc bound by the four cysteines is required for the DNA cleavage by PacI endonucleases. The H, DR, and CxxCN catalytic residues of 3M7K HNH endonuclease are generally conserved in the Opia proteins (e.g., Opia-19564_16, Figure 4.5C). However, the expected tyrosine residue precedes rather than follows the DR motif and its placement in the predicted structure is offset from that in 3M7K. The histidine active site residue is also slightly differently positioned (Figure 4.5C). These discrepancies may be attributed to uncertainties in protein folding. However, the positioning of histidine in the location typically occupied by tyrosine suggests its potential involvement in DNA cleavage, as occurs in other HNH endonucleases. Elsewhere, the predicted structures have large regions of positively charged surface, likely involved in DNA binding. These findings suggest that the Opia proteins share characteristics with PacI restriction endonucleases, yet they may represent a novel class of enzymes, likely with homing endonuclease function. (Figure 4.5B). The biochemically characterized PacI homodimer has a target recognition sequence of 5’-TTAATTAA-3’, and cleaves between the internal thymine residues. Pacl endonucleases rely on the absence of the recognition site elsewhere in the host genome. We could not determine the recognition sequence for the Opia Pacl-like homing restriction endonuclease, but apparently it was possible for different combinations of six variants to insert in the same region of a series of Opia genotypes.

We further reconstructed a circular, complete 26,349 bp genome (MGE-9917) for another circularized element that is targeted by three CRISPR spacers from Atabeya-1. MGE-9917 could not be classified as a virus or plasmid based on predicted protein functions. All MGE-9917 genes are encoded on the same strand. The genome features a CT-rich intergenic tandem repeat region with 6, 7, 8 or 9 units of 8 bp in length (variants identified using mapped reads). MGE-9917 encodes at least 14 proteins with transmembrane domains, two of which are 1,402 and 1,202 amino acids in length and lack related sequences in the NCBI database (Table S4). The genome contains a protein that combines an N-terminal ParB-like nuclease domain with a C-terminal S-adenosylmethionine-dependent methyltransferase domain. Additionally, MGE-9917 includes genes for a tyrosine recombinase, a transposase, and a Type IV methyl-directed restriction enzyme featuring an HNH motif. A family of related elements occurs in virtually all of the deep soil samples. One version differs due to the presence of a transposase that is related to those found in the Opia viruses.

Defense systems and epigenetic regulations in Atabeyarchaeia and Freyarchaeia

We used DefenseFinder (Tesson et al. 2022) and PADLOC (Payne et al. 2022) to identify ten defense systems in Freyarchaeia, five in Atabeyarchaeia-1, and six in Atabeyarchaeia-2. Freyarchaeia harbored at least four different defense system classes, including Type I-B, III-A, and III-D CRISPR-Cas systems, Type II restriction-modification systems, the Hachiman antiphage defense system, and the anti-phage system, DISARM (defense island system associated with restriction-modification), which has not previously been found in Asgard archaea (Figure S7-A, Table S5-S6). The DISARM system comprises *drmABC*, a methyltransferase (*drmMI*, N6 adenine-specific methyltransferase or *drmMII*, C5 cytosine-specific DNA methyltransferase), and *drmD* or *drmE*) (Ofir et al. 2018). The Freyarchaeia system includes *drmA*, *drmB, drmC, drmMII*, *drmE,* and drmD (helicase similar to the RNA polymerase (RNAP)-associated SWI2/SNF2 protein) that is present in, classifying this system as a DISARM class II (DISARM-II). Interestingly, *drmD* is a homolog typically found in DISARM class I. The DISARM methylase modifies host CCWGG motifs to distinguish its own DNA from foreign DNA. A specific conformation of the DrmAB complex (trigger loop) inhibits the complex to prevent an autoimmune response20. DrmA is responsible for DNA targeting in DISARM through multiple non-specific interactions with the DNA backbone (Bravo et al. 2022). By not requiring a specific sequence for DNA binding, DrmA distinguishes this defense system from other common restriction-modification systems, endowing DISARM with a broad spectrum of action against viruses (Tesson et al. 2022). Phylogenetic analysis of the helicase DrmA places the gene within the Euryarchaeia and Chloroflexota, suggesting that this system has been laterally transferred Figure S7-B. The Hachiman system is encoded by *hamA*, a hypothetical protein (DUF1837), and *hamB*, a helicase (pfam00271). The molecular mechanism of this system is still unknown (Doron et al. 2018). One potential hint is the presence of an ATP-dependent endonuclease is an OLD (overcoming lysogenization defect) upstream of the *hamAB* locus.

We used DNA polymerase kinetics from Pacific Biosciences (PacBio) metagenomic sequencing data and the Restriction Enzyme database (REBASE) to illuminate the DNA methylation patterns and methylases in the genomes of Freyarchaeia, Atabeyarchaeia-1, and Atabeyarchaeia-2. In the genome of Atabeyarchaeia-1, we identified 13 methylation sequence motifs, of which seven were directly linked to a specific methylase gene. Similarly, Atabeya-2 has 11 methylation motifs, 5 of which could be linked to a methylase. Freyarchaeia has only five detectable methylation motifs (Table S7A-B). Interestingly, one of those motifs is CCWGG, which has been characterized as a motif targeted by DISARM class II.

Atabeyarchaeia-1 and Atabeyarchaeia-2 both have 4-methylcytosine (m4C) and 6-methyladenosine (m6A) methylation. Atabeyarchaeia-1 Yucahu-i MGE encodes a Type IIG restriction-modification system that targets a m6A methylation motif and is the only candidate enzyme that could methylate the Atabeyarchaeia-1 genome. The Yucahu-i system is analogous to the MmeI family, which typically recognizes a 6-7 bp motif with adenine as the penultimate base22. The motif GYATGAG (m6A) was methylated at 66% of sites within the Atabeyarchaeia-1 genome, and could represent the active methylation motif of this RM system. In contrast, Freyarchaeia has 5 m4C motifs but no m6A methylation motifs.

Integrated mobile-like regions encode eukaryotic signature proteins

We identified two small GTPases in a region of the Atabeyarchaeia-1 genome that appears to be enriched in genes often associated with mobile genetic elements. This region is absent in some Atabeya-1 strains. The classification of these proteins as GTPases is supported by sequence and structural homology, as well as, structural predictions in comparison to reference eukaryotic sequences (Supplementary Text). Based on phylogenetic analysis, these proteins cluster with the unannotated Arf GTPases, ArfX, found in single-cell eukaryotes *Naegleria gruberi*, *Spironucleus salmonicida, Gefionella okellyi*, and Arfrp 1b in *Pygsuia biforma* (Figure S8).

4.4 Discussion

We identified chromosomally integrated and coexisting MGEs that replicate in Atabeyarchaeia by leveraging extensive sequencing of a series of soil samples where soil depth and biogeochemical conditions select for different strain variant populations. By examining read mappings across multiple samples from the same environment, we established a larger repertoire of MGEs than could be found from the analysis of any single metagenome. Integrated elements and coexisting circularized MGEs range from 2.5 to 40 Kb in length; all complete MGEs were circular and at least some replicate bidirectionally. We could not confidently identify MGEs by changes in read mapping abundances in the Freyarchaeia genome. This might indicate either stable integration into the Freyarchaeia genome across the entire population studied (thus excision sites and coexisting versions of circular elements were not detected by our methods) or that , the degree of association with MGEs may vary dramatically between Asgard archaeal lineages.

The presence of coexisting integrated and free, circularized MGEs, likely mostly plasmids, as well as variation in copy number of circularized elements and in the fraction of cells with integrated elements suggest regular movement of MGEs into and out of the Atabeyarchaeia chromosomes. Insertion/excision and variation in copy number may enable Atabeyarchaeia to respond to changes in their environment. For example, MGEs may behave symbiotically, and increase in MGE copy number (thus gene content) serves as a response to increased pressure from other MGE (Krupovic et al., 2019). The tiny mini-Yucahu indicates another layer of genomic variability, as just this portion of the host MGE can excise.

Interestingly, the attachment motifs for Yucahu-i plasmid-like Atabeya-1 MGE and for a circular, unclassified and essentially unrelated MGE linked to Atabeya-2 are exactly the same, implying that the very different integrases of each (25% aa ID) recognize and cut at the same motif. Protein sequence divergence may enable host chromosome specificity, yet the active site apparently evolved to target the same motif.

The novel cluster of genomically similar Opia viruses of Atabeya-2 and 2’ encode sequential cysteine-rich proteins inferred to have nuclease activity due to their distant homology to PacI. These occur combinatorial patterns (Figure 4.5B) that may have arisen via recent recombination events in which these putative endonucleases may have played a role. The multiple variants might form heterodimers rather than the normal homodimers expected for Pacl, possibly extending target recognition. The results suggest the importance of diverse nuclease activity for these viruses.

The Opia virus proteomes exhibit similarities (e.g., capsid, tube and tail proteins)to those of tailed viruses, which commonly replicate in bacterial and archaeal hosts from hypersaline environments (Senčilo & Roine, 2014). They are quite distinct from those of eukaryotic viruses, supporting the suggestion that, despite the evolutionary relationship between Asgard archaea and eukaryotes, their viruses display no obvious evolutionary relationships (Medvedeva et al., 2022; Rambo et al., 2022; Tamarit et al., 2022).

Genome context (dominated by MGE-associated genes) and apparent excision of the region encoding the GTPases from some strain genotypes supports the inference that some Atabeyarchaeia MGEs encode eukaryotic signature proteins. ARF GTPases are involved in membrane trafficking in eukaryotes and have been previously described as ESPs in Asgard archaea (Eme et al., 2023; Spang et al., 2015). The presence of these GTPases on putative MGEs provides the first indication that increase in cellular complexity could be associated with the transfer of eukaryotic signature proteins via MGE (Figure S8, Figure S9, Table S8).

Our results suggest several examples of genes integrated into Atabeyarchaeia genomes or in their coexisting MGEs that have nearest homologs in the bacterial domain (e.g., ParB, DrmA, and Type IIG restriction-modification enzymes). These findings are consistent with recent work on cross-domain gene transfer via movement of integrons associated with diverse MGEs (Ghaly et al., 2022) and extend earlier work inferring the acquisition of archaeal genes by bacteria (e.g., (Hug et al., 2013)). As we discover new Asgard archaea from genome-resolved metagenomes, we can expect to find further parallels between bacterial and archaeal immune systems. These associations could have implications for the evolution of eukaryotic immune systems (Wein & Sorek, 2022).

Type IIG restriction-modification systems, to our knowledge, have not previously been associated with archaeal MGEs. The observation that the only methylase seemingly able to methylate at the host genome's well-represented m6A motif is carried by Yucahu-i, and that it is transcriptionally active, suggests that the Yucahu-i Type IIG plays a significant role in host genome epigenetic modification. This MGE-encoded system may protect its host Atabeyarchaeia against infection by other MGEs. This behavior aligns with the emerging perspective that defense systems themselves can serve as mobile genetic elements (Rocha & Bikard, 2022; Wu et al., 2022).

To our knowledge, these are the first metagenome-derived Asgard archaeal complete genomes for which methylation patterns have been reported. PacBio sequences corresponding to these complete, manually curated genomes (Valentin-Alvarado et al. 2023) were used to infer the methylation motifs and to determine the fraction of sites that were methylated, as well as the methylation patterns of their newly reported MGEs. Using REBASE, which features all biochemically characterized methylases, it was possible to link methylated sites with likely methylases encoded on both genomes and MGEs. Relatively little is known about genome methylation in archaea, especially in Asgard archaea (Anton & Roberts, 2021). These genomes and their methylation motif data presented here provide a starting point for detailed biochemical studies to expand the known inventory of archaeal methylases. The higher number of methylation motifs in Atabeyarchaeia compared to the Freyarchaeia genome could be an evolutionary response to the larger inventory of mobile genetic elements associated with Atabeyarcheia.

Our results brought to light new MGEs, defense systems, and epigenetic patterns in these soil-associated Asgard archaea. The description and annotation of the first complete Asgard plasmids opens a route toward the development of tools for genetic manipulation of these organisms.

4.5 Conclusion

We leveraged the read diversity inherent to population genomic data, long-read sequencing, methylation pattern analysis, comparative genomics and functional and structural prediction to explore integrated and coexisting genetic elements of one group of Asgard archaea. These analyses brought to light an extensive landscape of mobile genetic elements (MGEs) that associate with Atabeyarchaeia, including viruses, plasmids and as yet unclassified entities. The excision, insertion and changes in copy number of these MGEs may enable adaptation to changing conditions, have contributed evolution, and possibly to the acquisition and spread of genes linked to the origin of cellular complexity. The availability of MGEs that could be adapted for delivery of genome editing tools in a community context (Rubin et al. 2022) may pave the way for genetic manipulation of these archaea.

4.6 Figures

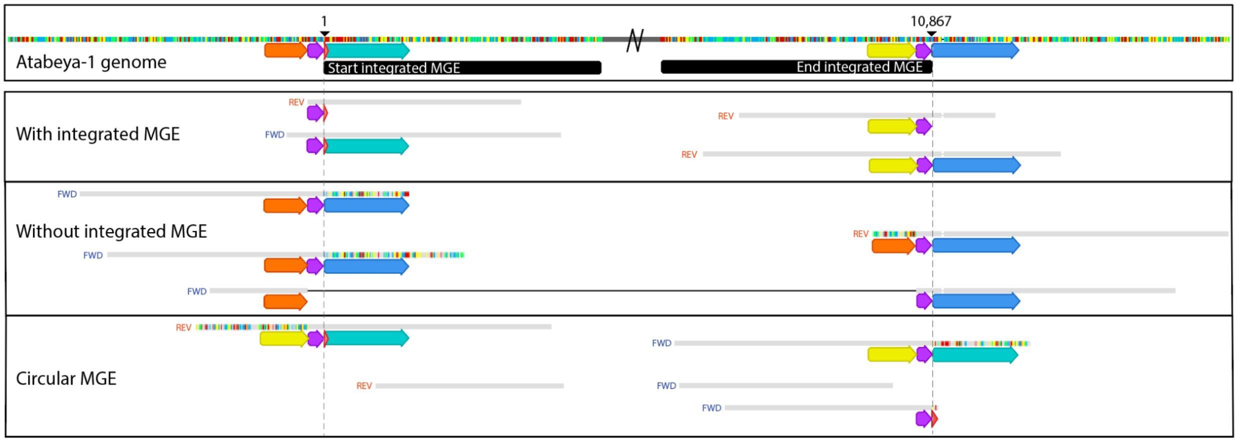


Figure 4.1 Read mapping to the reference genome provides evidence for integration and excision, illustrated for the case of one MGE. The central region of the integrated sequence of Yucahu-i (between black bars) has been deleted to focus on details of reads mapped to the start and end of the region. Read sequences that match the genome sequences are shown as gray bars; small vertical colored bars adjacent to read portions in agreement with the reference indicate bases that disagree with the reference. Arrows with the same color have the same nucleotide sequence. In the panel demonstrating that some cells that lack the integrated MGE, one read has been split (the black line links the two parts of a single read) to illustrate agreement with the flanking sequence at both ends of the integrated region.

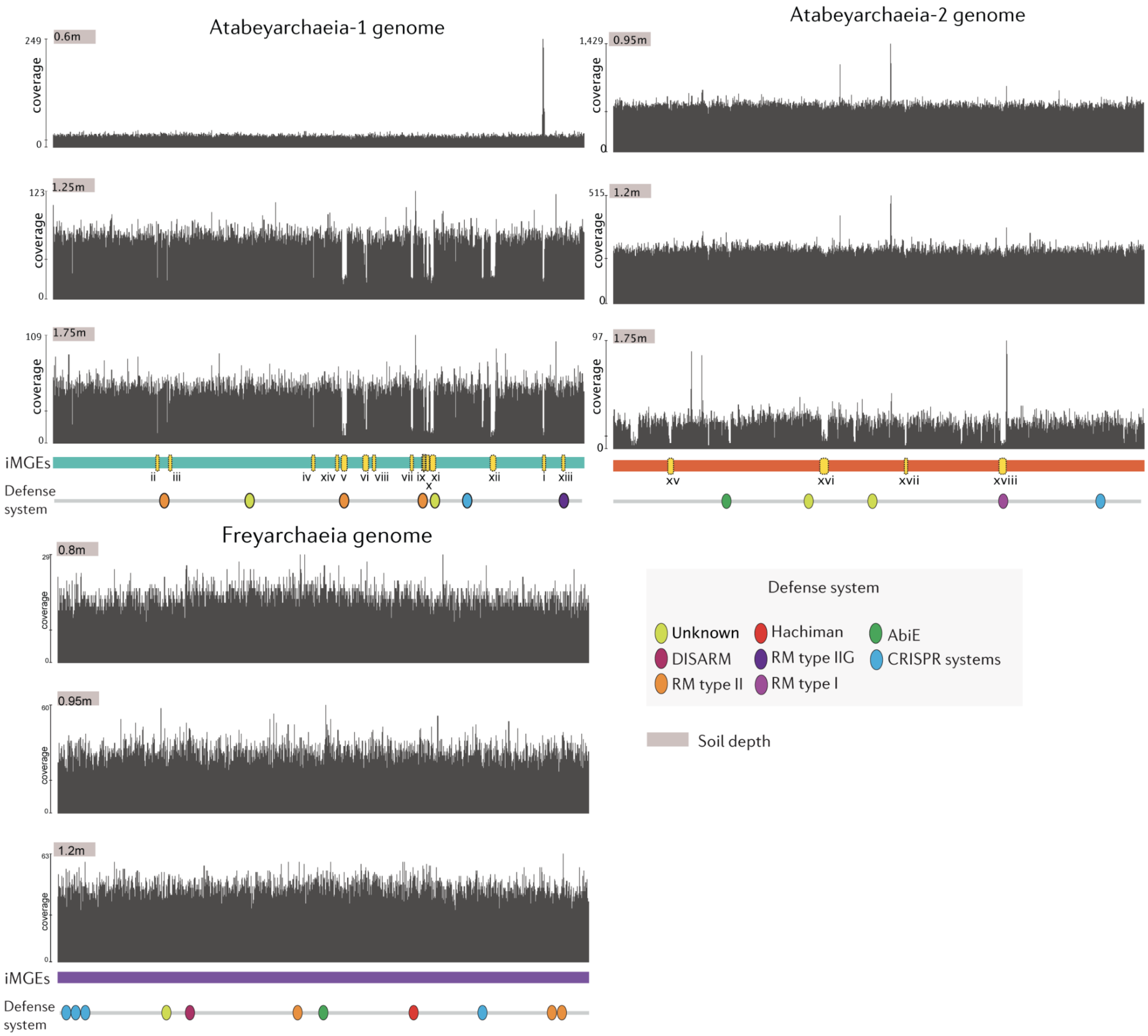


Figure 4.2 Chromosomally integrated genetic elements and defense systems in soil Asgard Archaea genomes. Each panel depicts the coverage across each complete genome, as determined by mapping to metagenome reads derived from three different soil depth profiles. Regions exhibiting low coverage suggest strain variations associated with specific soil depths and may indicate the presence of integrated genetic elements in only a subset of cells. Notably, the Freyarchaeia genome exhibits even coverage using reads from all sampling depths, with no discernible integrated mobile genetic elements identified. Some of the low-coverage regions are not labeled as potential MGEs, these regions are strain variants with sequences so divergent that read mapping is precluded. Oval symbols indicate predicted defense systems.

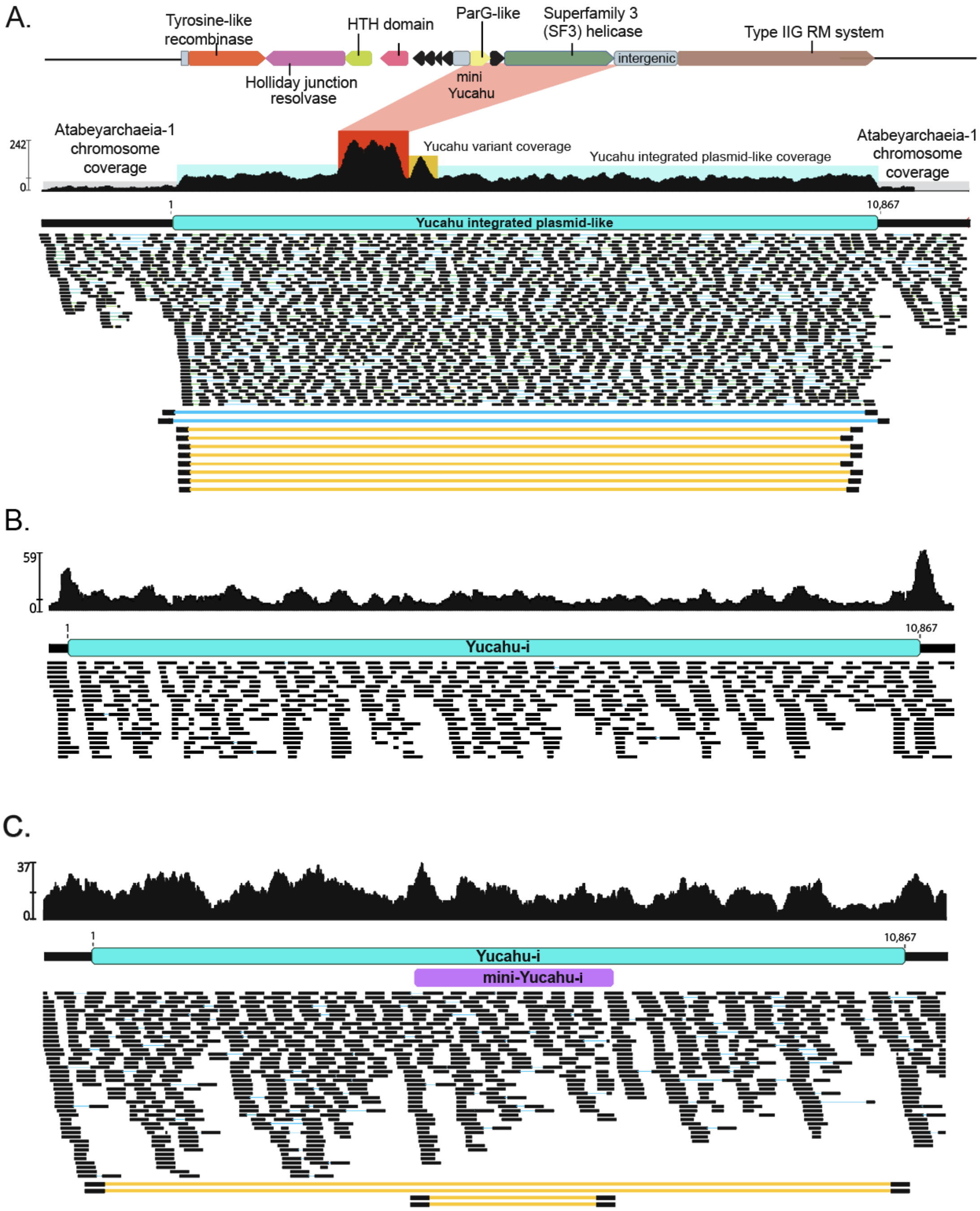


Figure 4.3 Integration and excision of Yucahu. A) For the 60 cm sample, elevated coverage and paired reads indicate that Yucahu-i is integrated into the genome, excised from some genomes (blue lines) and coexists in circularized form (yellow lines). The red box indicates elevated coverage from a related gene from another genome. B) For the 165 cm sample, low coverage over the MGE and read sequence discrepancies indicate that most cells in this sample lack the MGE. C) For the 70 cm sample, coverage and paired read information indicate that Yucahu-i is integrated into essentially all cells. The circularized Yucahu-i is present but rare. Paired reads pointing out internal to the MGE indicate that a 2,644 bp element has integrated into the plasmid and coexists in circularized form.

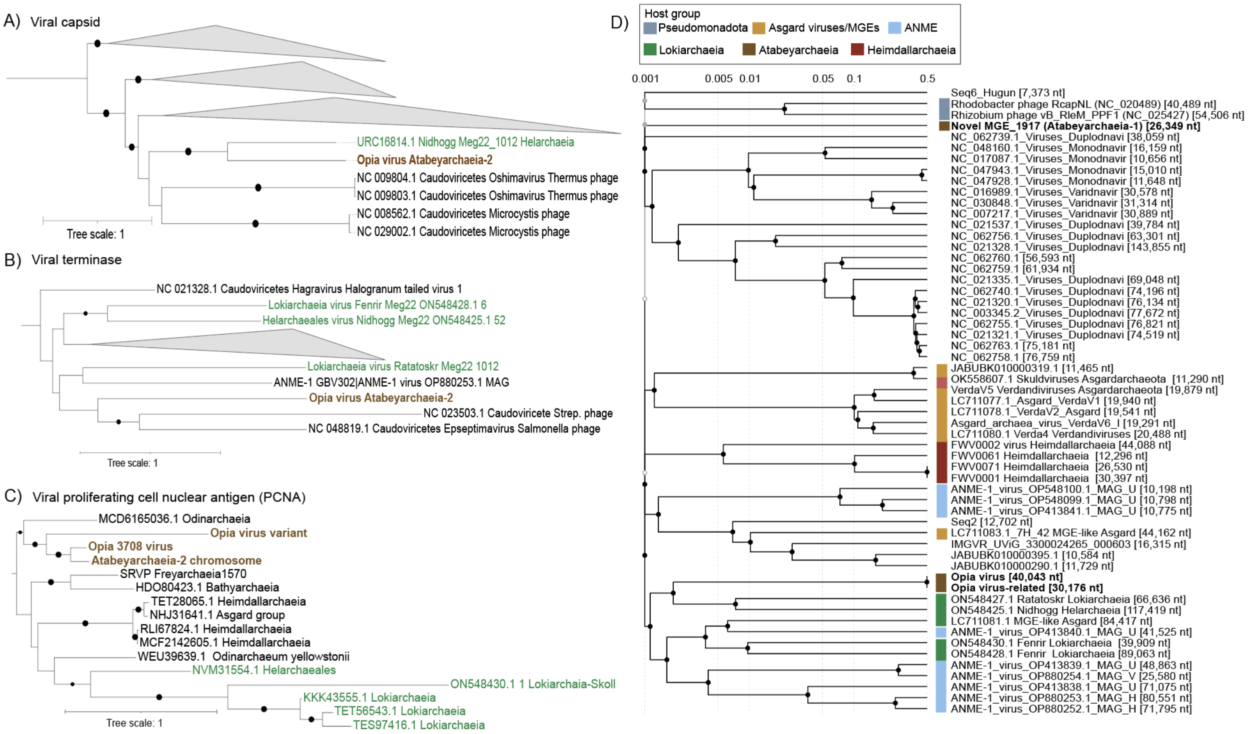

A) Viral capsid
URC16814.1 Nidhogg Meg22_1012 Helarchaeia
Opia virus Atabeyarchaeia-2
NC 009804.1 Caudoviricetes Oshimavirus Thermus phage
NC 009803.1 Caudoviricetes Oshimavirus Thermus phage
NC 008562.1 Caudoviricetes Microcystis phage
NC 029002.1 Caudoviricetes Microcystis phage
Tree scale: 1
B) Viral terminase
NC 021328.1 Caudoviricetes Hagravirus Halogranum tailed virus 1
Lokiarchaeia virus Fenrir Meg22 ON548428.1 6
Helarchaeales virus Nidhogg Meg22 ON548425.1 52
Lokiarchaeia virus Ratatoskr Meg22 1012
ANME-1 GBV302|ANME-1 virus OP880253.1 MAG
Opia virus Atabeyarchaeia-2
NC 023503.1 Caudoviricete Strep. phage
NC 048819.1 Caudoviricetes Epseptimavirus Salmonella phage
Tree scale: 1
C) Viral proliferating cell nuclear antigen (PCNA)
MCD6165036.1 Odinarchaeia
Opia virus variant
Opia 3708 virus
Atabeyarchaeia-2 chromosome
SRVP Freyarchaeia1570
HDO80423.1 Bathyarchaeia
TET28065.1 Heimdallarchaeia
NHJ31641.1 Asgard group
RLI67824.1 Heimdallarchaeia
MCF2142605.1 Heimdallarchaeia
WEU39639.1 Odinarchaeum yellowstonii
NVM31554.1 Helarchaeales
ON548430.1 1 Lokiarchaia-Skoll
KKK43555.1 Lokiarchaeia
TET56543.1 Lokiarchaeia
TES97416.1 Lokiarchaeia
Tree scale: 1
D)
Host group
Pseudomonadota
Asgard viruses/MGEs
ANME
Lokiarchaeia
Atabeyarchaeia
Heimdallarchaeia
0.001
0.005
0.01
0.05
0.1
0.5
Seq6_Hugun [7,373 nt]
Rhodobacter phage RcapNL (NC_020489) [40,489 nt]
Rhizobium phage vB_RleM_PPF1 (NC_025427) [54,506 nt]
Novel MGE_1917 (Atabeyarchaeia-1) [26,349 nt]
NC_062739.1_Viruses_Duplodnavi [38,059 nt]
NC_048160.1_Viruses_Monodnavir [16,159 nt]
NC_017087.1_Viruses_Monodnavir [10,656 nt]
NC_047943.1_Viruses_Monodnavir [15,010 nt]
NC_047928.1_Viruses_Monodnavir [11,648 nt]
NC_016989.1_Viruses_Varidnavir [30,578 nt]
NC_030848.1_Viruses_Varidnavir [31,314 nt]
NC_007217.1_Viruses_Varidnavir [30,889 nt]
NC_021537.1_Viruses_Duplodnavi [39,784 nt]
NC_062756.1_Viruses_Duplodnavi [63,301 nt]
NC_021328.1_Viruses_Duplodnavi [143,855 nt]
NC_062760.1 [56,593 nt]
NC_062759.1 [61,934 nt]
NC_021335.1_Viruses_Duplodnavi [69,048 nt]
NC_062740.1_Viruses_Duplodnavi [74,196 nt]
NC_021320.1_Viruses_Duplodnavi [76,134 nt]
NC_003345.2_Viruses_Duplodnavi [77,672 nt]
NC_062755.1_Viruses_Duplodnavi [76,821 nt]
NC_021321.1_Viruses_Duplodnavi [74,519 nt]
NC_062763.1 [75,181 nt]
NC_062758.1 [76,759 nt]
JABUBK010000319.1 [11,465 nt]
OK558607.1 Skuldviruses Asgardarchaeota [11,290 nt]
VerdaV5 Verdandiviruses Asgardarchaeota [19,879 nt]
LC711077.1_Asgard_VerdaV1 [19,940 nt]
LC711078.1_VerdaV2_Asgard [19,541 nt]
Asgard_archaea_virus_VerdaV6_I [19,291 nt]
LC711080.1 Verda4 Verdandiviruses [20,488 nt]
FWV0002 virus Heimdallarchaeia [44,088 nt]
FWV0061 Heimdallarchaeia [12,296 nt]
FWV0071 Heimdallarchaeia [26,530 nt]
FWV0001 Heimdallarchaeia [30,397 nt]
ANME-1_virus_OP548100.1_MAG_U [10,198 nt]
ANME-1_virus_OP548099.1_MAG_U [10,798 nt]
ANME-1_virus_OP413841.1_MAG_U [10,775 nt]
Seq2 [12,702 nt]
LC711083.1_7H_42 MGE-like Asgard [44,162 nt]
IMGVR_UViG_3300024265_000603 [16,315 nt]
JABUBK010000395.1 [10,584 nt]
JABUBK010000290.1 [11,729 nt]
Opia virus [40,043 nt]
Opia virus-related [30,176 nt]
ON548427.1 Ratatoskr Lokiarchaeia [66,636 nt]
ON548425.1 Nidhogg Helarchaeia [117,419 nt]
LC711081.1 MGE-like Asgard [84,417 nt]
ANME-1_virus_OP413840.1_MAG_U [41,525 nt]
ON548430.1 Fenrir Lokiarchaeia [39,909 nt]
ON548428.1 Fenrir Lokiarchaeia [89,063 nt]
ANME-1_virus_OP413839.1_MAG_U [48,863 nt]
ANME-1_virus_OP880254.1_MAG_V [25,580 nt]
ANME-1_virus_OP413838.1_MAG_U [71,075 nt]
ANME-1_virus_OP880253.1_MAG_H [80,551 nt]
ANME-1_virus_OP880252.1_MAG_H [71,795 nt]

Figure 4.4 Phylogenetic placement of hallmark structural viral proteins from Opia and whole proteome-based similarity of MGE from Atabeyarchaeia and other Asgardarchaeota. A) Viral capsid. B) Viral terminase. C) viral proliferating cell nuclear antigen (PCNA) and D) whole proteome-based similarity of MGE from Atabeyarchaeia and other Asgardarchaeota. Boostraps are represented by black circules >80%.

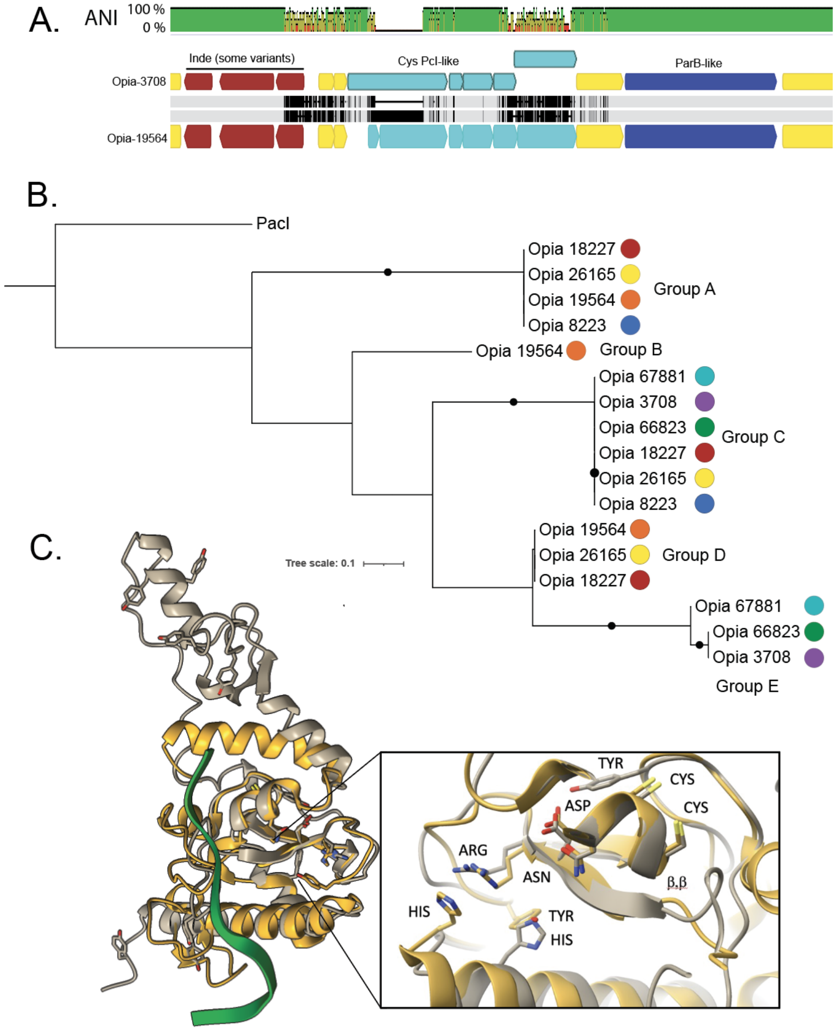


Figure 4.5 Sequence variation in the cysteine-rich proteins of Opia viruses and comparison to PacI, a rare-cutting HNH restriction endonuclease. A) The aligned ~5.6 kbp variable region of two Opia viral genotypes. Light gray bars indicate perfect nucleotide identity and thin vertical black lines indicate SNPs. The three genes labeled in brown are present in both Opia-3708 (top) and Opia-19564 (bottom) but are absent in some other genotypes. Light blue genes are cysteine-rich (For 3708L, 10, 5, 9, 5, 10 and for Opia-19564, 3, 14, 5, 9, 4, 10 cysteines per protein). Dark blue genes encode ParB-like proteins that are very divergent between some genotypes (Opia-3708 vs. Opia-66823, 67881). Before the 3-gene indel and after the ParB-like gene, the ~35 kb regions of all genomes are essentially identical. B) Phylogenetic tree including the 17 larger cysteine-rich proteins from the variable regions of 7 Opia genomes (numbers represent genome names; for context, see supplementary data. Proteins with identical sequences occur in five different combinations across the genotypes. C) Comparison of the active site region of PacI (with 9 cysteines) and the structure of Opia-19564 protein 16 (silver, with 10 cysteines). Active site residues of PDB 3m7k (gold) are displayed based on Figure 1 of Shen et al. (2010). The critical ββα-metal motifs are well aligned. Tyrosine and histidine residues exist in proximity to the active site, although their locations differ somewhat, possibly due to fold inaccuracy.

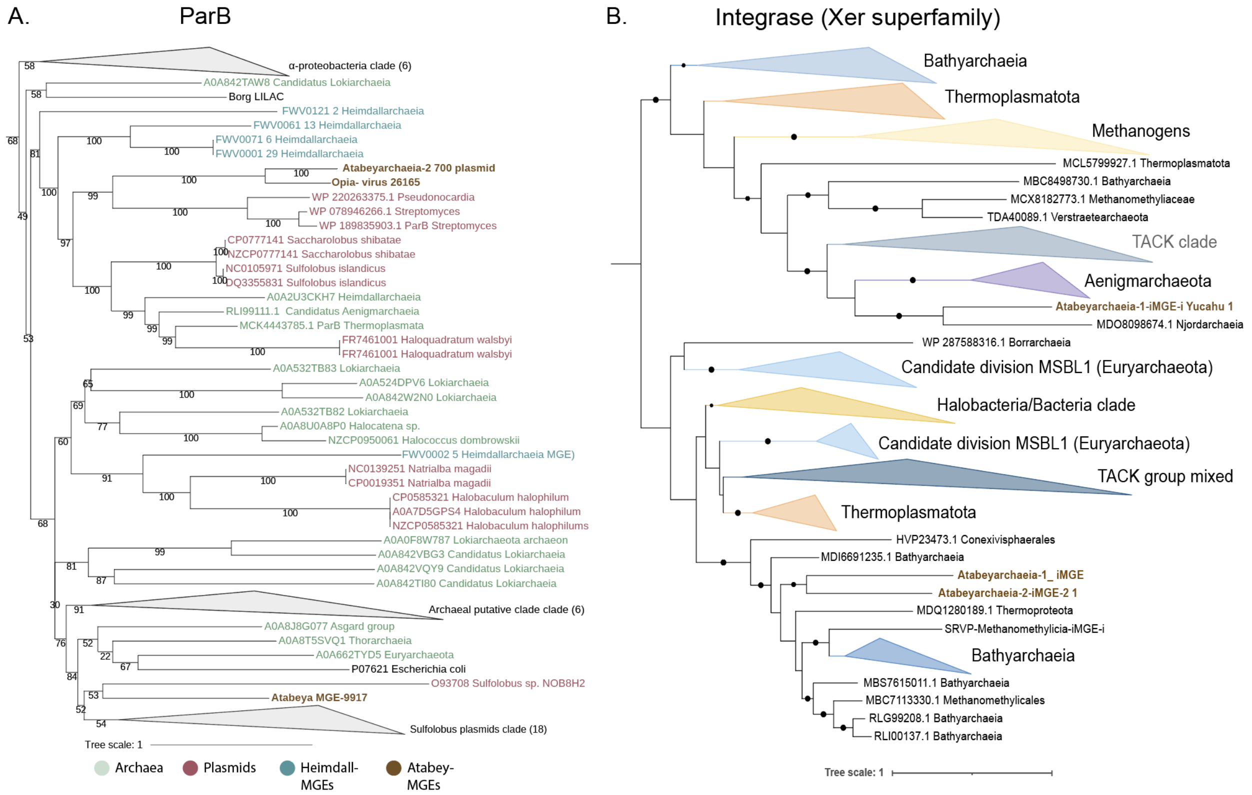

A.
ParB
58
α-proteobacteria clade (6)
A0A842TAW8 Candidatus Lokiarchaeia
58
Borg LILAC
FWV0121 2 Heimdallarchaeia
FWV0061 13 Heimdallarchaeia
68
100
FWV0071 6 Heimdallarchaeia
81
100
FWV0001 29 Heimdallarchaeia
Atabeyarchaeia-2 700 plasmid
100
100
Opia- virus 26165
100
99
WP 220263375.1 Pseudonocardia
100
WP 078946266.1 Streptomyces
49
100
WP 189835903.1 ParB Streptomyces
97
CP0777141 Saccharolobus shibatae
100
NZCP0777141 Saccharolobus shibatae
100
NC0105971 Sulfolobus islandicus
100
DQ3355831 Sulfolobus islandicus
100
A0A2U3CKH7 Heimdallarchaeia
RLI99111.1 Candidatus Aenigmarchaeia
99
MCK4443785.1 ParB Thermoplasmata
99
99
FR7461001 Haloquadratum walsbyi
53
100
FR7461001 Haloquadratum walsbyi
A0A532TB83 Lokiarchaeia
65
A0A524DPV6 Lokiarchaeia
100
A0A842W2N0 Lokiarchaeia
69
A0A532TB82 Lokiarchaeia
77
A0A8U0A8P0 Halocatena sp.
60
100
NZCP0950061 Halococcus dombrowskii
FWV0002 5 Heimdallarchaeia MGE)
91
NC0139251 Natrialba magadii
100
CP0019351 Natrialba magadii
100
CP0585321 Halobaculum halophilum
100
A0A7D5GPS4 Halobaculum halophilum
68
NZCP0585321 Halobaculum halophilums
A0A0F8W787 Lokiarchaeota archaeon
99
A0A842VBG3 Candidatus Lokiarchaeia
81
A0A842VQY9 Candidatus Lokiarchaeia
87
A0A842TI80 Candidatus Lokiarchaeia
30
91
Archaeal putative clade clade (6)
A0A8J8G077 Asgard group
76
52
A0A8T5SVQ1 Thorarchaeia
22
A0A662TYD5 Euryarchaeota
67
P07621 Escherichia coli
84
O93708 Sulfolobus sp. NOB8H2
53
Atabeya MGE-9917
52
54
Sulfolobus plasmids clade (18)
Tree scale: 1
Archaea
Plasmids
Heimdall-MGEs
Atabey-MGEs
B.
Integrase (Xer superfamily)
Bathyarchaeia
Thermoplasmatota
Methanogens
MCL5799927.1 Thermoplasmatota
MBC8498730.1 Bathyarchaeia
MCX8182773.1 Methanomethyliaceae
TDA40089.1 Verstraetearchaeota
TACK clade
Aenigmarchaeota
Atabeyarchaeia-1-iMGE-i Yucahu 1
MDO8098674.1 Njordarchaeia
WP 287588316.1 Borrarchaeia
Candidate division MSBL1 (Euryarchaeota)
Halobacteria/Bacteria clade
Candidate division MSBL1 (Euryarchaeota)
TACK group mixed
Thermoplasmatota
HVP23473.1 Conexivisphaerales
MDI6691235.1 Bathyarchaeia
Atabeyarchaeia-1_ iMGE
Atabeyarchaeia-2-iMGE-2 1
MDQ1280189.1 Thermoproteota
SRVP-Methanomethylicia-iMGE-i
Bathyarchaeia
MBS7615011.1 Bathyarchaeia
MBC7113330.1 Methanomethylicales
RLG99208.1 Bathyarchaeia
RLI00137.1 Bathyarchaeia
Tree scale: 1

Figure S1 Phylogenetic analyses of ParB and tyrosine-recombinase-like proteins across archaeal mobile genetic elements. A) Phylogenetic tree for ParB and ParB-like Proteins. Analysis incorporates reference sequences of *E. coli* ParB (Pfam08775) and *Sulfolobus* conjugative plasmid ParB (NOB8H2), alongside representative ParB-like domains from select archaeal and bacterial genomes. Selection criteria for inclusion involved a rigorous screening of the top 25 protein hits from the NCBI database. Sequence alignments were conducted using MAFFT (v7.310) in 'auto' mode, with subsequent optimization by trimAl (v1.4.rev15) applying a 0.7 gap threshold. Phylogenetic trees were initially inferred using IQ-TREE (v1.6.1) under the LG+FO+R model. This segment details B) Phylogeny of tyrosine-recombinase-like proteins within integrated genetic elements of Atabeyarchaeia-1 and Atabeyarchaeia-2. The analysis selected the top 25 protein hits from the NCBI database, aligned with MAFFT (version 7.310) on 'auto' setting, and refined the alignment with trimAl (version 1.4.rev15) using a 0.7 gap threshold. The phylogenetic framework was constructed using IQ-TREE (version 1.6.1), adopting the LG+FO+R model.

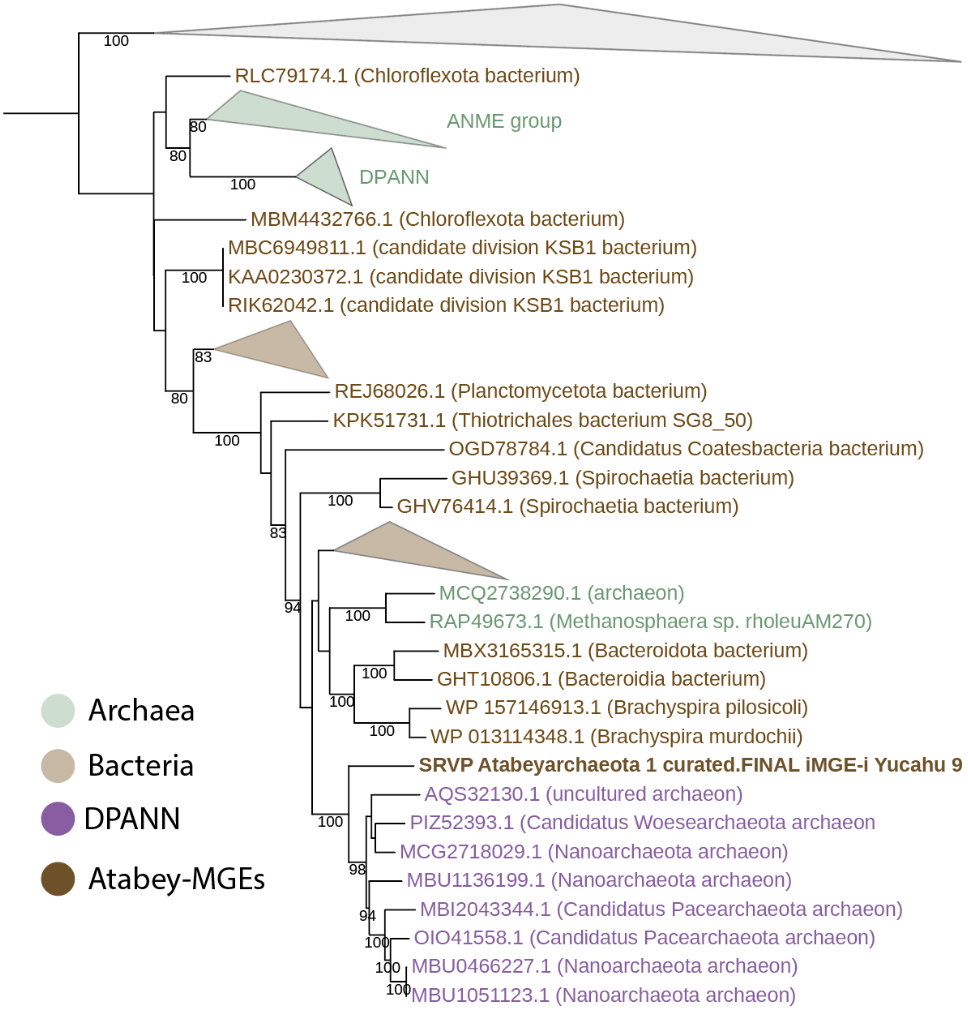

100
RLC79174.1 (Chloroflexota bacterium)
80
ANME group
80
100
DPANN
MBM4432766.1 (Chloroflexota bacterium)
MBC6949811.1 (candidate division KSB1 bacterium)
100
KAA0230372.1 (candidate division KSB1 bacterium)
RIK62042.1 (candidate division KSB1 bacterium)
83
80
REJ68026.1 (Planctomycetota bacterium)
KPK51731.1 (Thiotrichales bacterium SG8_50)
100
OGD78784.1 (Candidatus Coatesbacteria bacterium)
GHU39369.1 (Spirochaetia bacterium)
100
GHV76414.1 (Spirochaetia bacterium)
83
94
MCQ2738290.1 (archaeon)
100
RAP49673.1 (Methanosphaera sp. rholeuAM270)
MBX3165315.1 (Bacteroidota bacterium)
100
GHT10806.1 (Bacteroidia bacterium)
100
WP 157146913.1 (Brachyspira pilosicoli)
100
WP 013114348.1 (Brachyspira murdochii)
SRVP Atabeyarchaeota 1 curated.FINAL iMGE-i Yucahu 9
AQS32130.1 (uncultured archaeon)
100
PIZ52393.1 (Candidatus Woesearchaeota archaeon
MCG2718029.1 (Nanoarchaeota archaeon)
98
MBU1136199.1 (Nanoarchaeota archaeon)
94
MBI2043344.1 (Candidatus Pacearchaeota archaeon)
100
OIO41558.1 (Candidatus Pacearchaeota archaeon)
100
MBU0466227.1 (Nanoarchaeota archaeon)
100
MBU1051123.1 (Nanoarchaeota archaeon)
● Archaea
● Bacteria
● DPANN
● Atabey-MGEs

Figure S2 Maximum likelihood phylogeny of Type II-G restriction-modification (IIG RM) protein fusion that combines endonuclease and methyltransferase present in Yucahy MGE. The phylogenetic tree was initially constructed with IQ-TREE (version 1.6.1) employing the LG+FO+R model for the first iterations. Subsequent refinement involved several rounds of manual branch checking to ensure the accuracy and reliability of the phylogenetic relationships depicted.

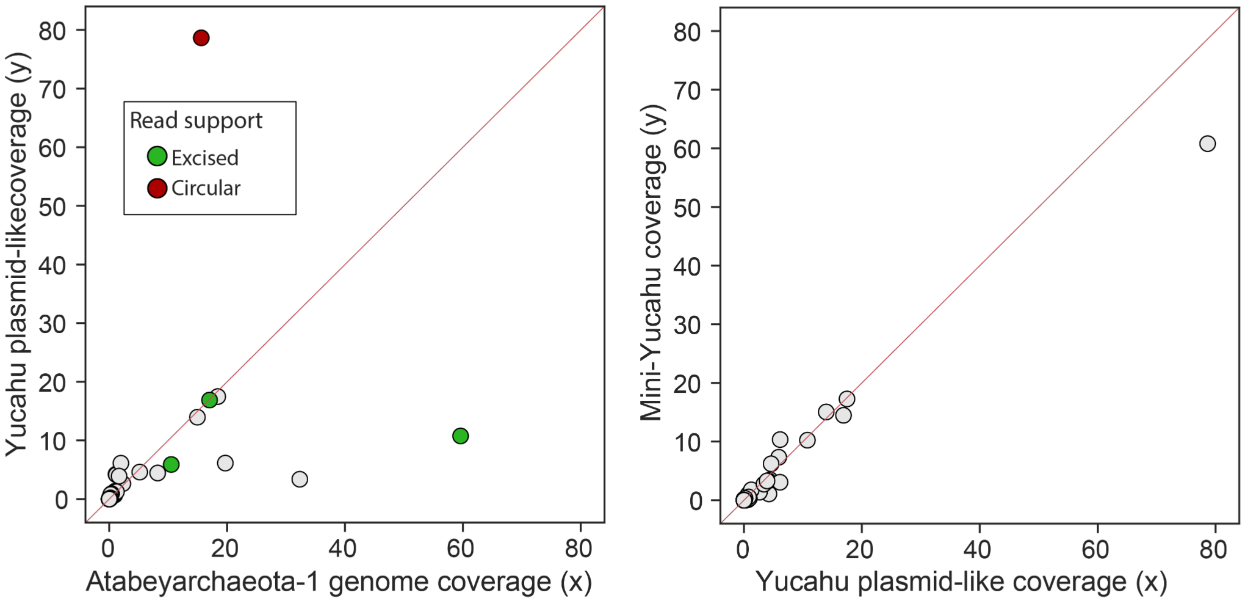


Figure S3 Cross-sample Atabeyarchaeia-1 plasmid integration. Correlations of the genome coverages between integrated plasmid Yucahu and Atabeyarchaeia and between the Yucahu plasmid genome and the mini-Yucahu element.

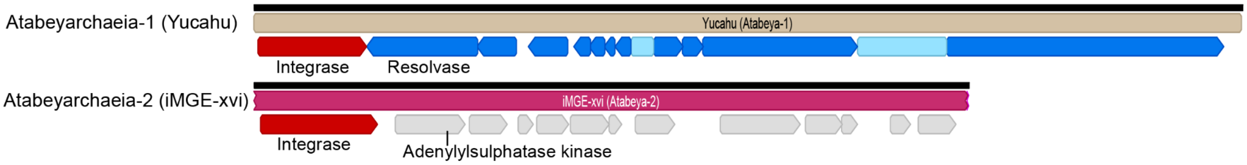

Atabeyarchaeia-1 (Yucahu)
Yucahu (Atabeya-1)
Integrase
Resolvase
Atabeyarchaeia-2 (iMGE-xvi)
iMGE-xvi (Atabeya-2)
Integrase
Adenylylsulphatase kinase

Figure S4 Comparison of gene content of Yucahu-i (Atabeyarchaeia-1), top. and iMGE-xvi **(Atabeyarchaeia-2), bottom.** Annotated proteins of Yucahu-i include a tyrosine recombinase and a protein with similarity to adenylylsulfate kinases. The iMGE-xvi has 12 open reading frames, most encoding hypothetical proteins or proteins of unknown function. Three proteins are predicted to have 5 - 8 transmembrane domains. One is a putative membrane-bound serine protease of the ClpP class involved in the proteolysis of misfolded and defective proteins (Moreno-Cinos et al. 2019)**.**

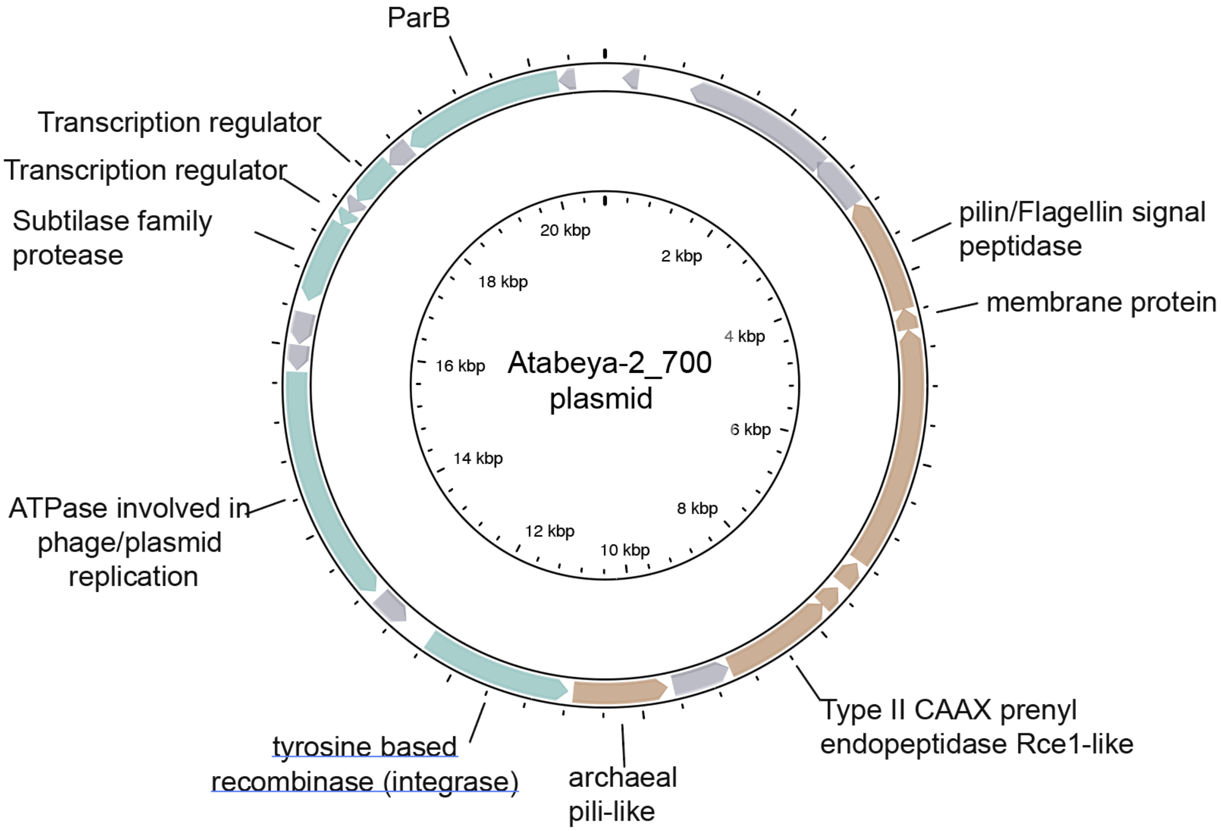

ParB
Transcription regulator
Transcription regulator
Subtilase family protease
ATPase involved in phage/plasmid replication
tyrosine based recombinase (integrase)
archaeal pili-like
Type II CAAX prenyl endopeptidase Rce1-like
membrane protein
pilin/Flagellin signal peptidase
Atabeya-2_700 plasmid
20 kbp
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp

Figure S5 Genome diagram of Atabeyarchaeia-2 putative plasmid scaffold_700. The plasmid has 23 open reading frames and gray-colored genes are hypothetical. Genes colored brown have transmembrane domains and are predicted to be extracellular.

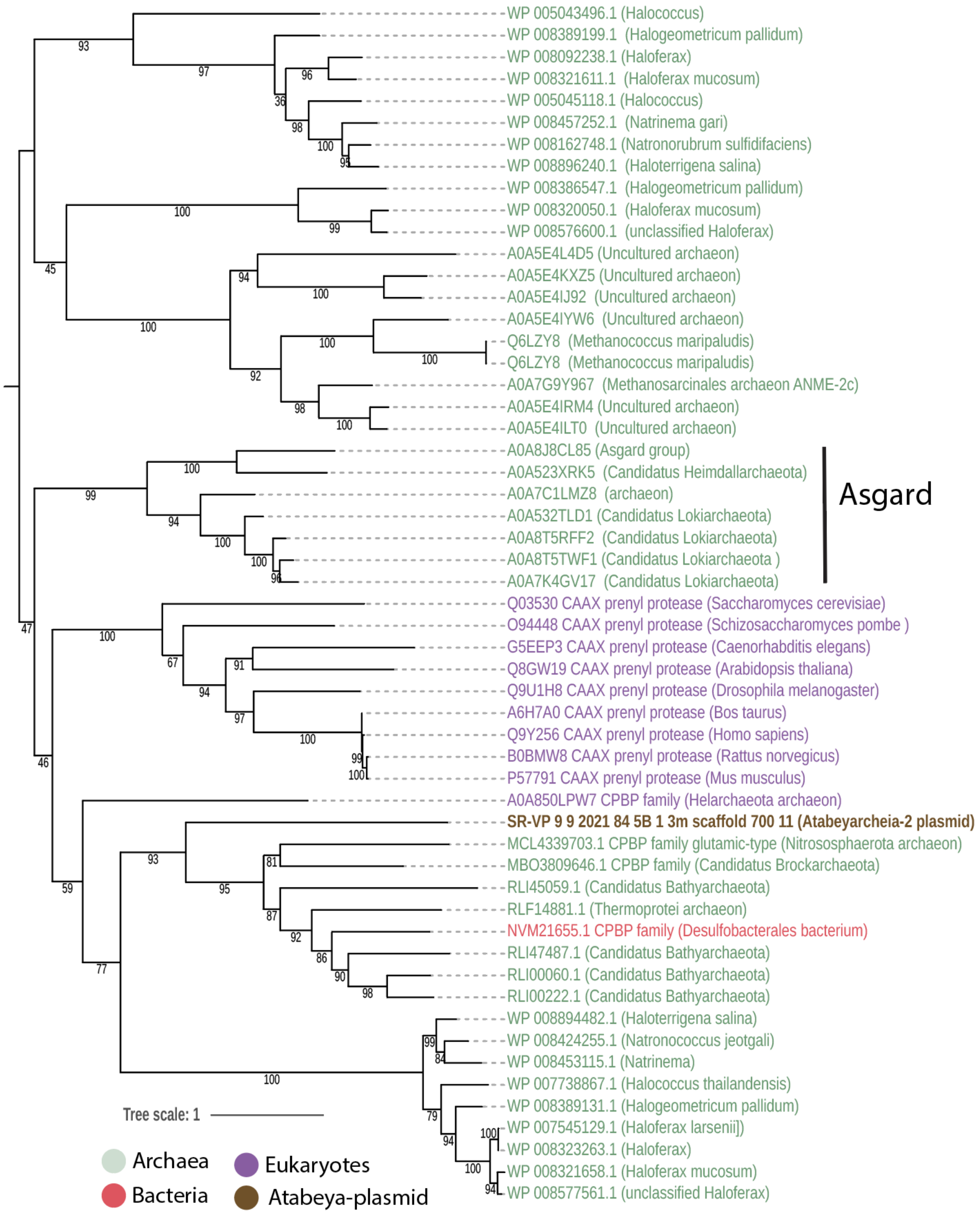

WP 005043496.1 (Halococcus)
WP 008389199.1 (Halogeometricum pallidum)
WP 008092238.1 (Haloferax)
WP 008321611.1 (Haloferax mucosum)
WP 005045118.1 (Halococcus)
WP 008457252.1 (Natrinema gari)
WP 008162748.1 (Natronorubrum sulfidifaciens)
WP 008896240.1 (Haloterrigena salina)
WP 008386547.1 (Halogeometricum pallidum)
WP 008320050.1 (Haloferax mucosum)
WP 008576600.1 (unclassified Haloferax)
A0A5E4L4D5 (Uncultured archaeon)
A0A5E4KXZ5 (Uncultured archaeon)
A0A5E4IJ92 (Uncultured archaeon)
A0A5E4IYW6 (Uncultured archaeon)
Q6LZY8 (Methanococcus maripaludis)
Q6LZY8 (Methanococcus maripaludis)
A0A7G9Y967 (Methanosarcinales archaeon ANME-2c)
A0A5E4IRM4 (Uncultured archaeon)
A0A5E4ILT0 (Uncultured archaeon)
A0A8J8CL85 (Asgard group)
A0A523XRK5 (Candidatus Heimdallarchaeota)
A0A7C1LMZ8 (archaeon)
A0A532TLD1 (Candidatus Lokiarchaeota)
A0A8T5RFF2 (Candidatus Lokiarchaeota)
A0A8T5TWF1 (Candidatus Lokiarchaeota)
A0A7K4GV17 (Candidatus Lokiarchaeota)
Asgard
Q03530 CAAX prenyl protease (Saccharomyces cerevisiae)
O94448 CAAX prenyl protease (Schizosaccharomyces pombe)
G5EEP3 CAAX prenyl protease (Caenorhabditis elegans)
Q8GW19 CAAX prenyl protease (Arabidopsis thaliana)
Q9U1H8 CAAX prenyl protease (Drosophila melanogaster)
A6H7A0 CAAX prenyl protease (Bos taurus)
Q9Y256 CAAX prenyl protease (Homo sapiens)
B0BMW8 CAAX prenyl protease (Rattus norvegicus)
P57791 CAAX prenyl protease (Mus musculus)
A0A850LPW7 CPBP family (Helarchaeota archaeon)
SR-VP 9 9 2021 84 5B 1 3m scaffold 700 11 (Atabeyarcheia-2 plasmid)
MCL4339703.1 CPBP family glutamic-type (Nitrososphaerota archaeon)
MBO3809646.1 CPBP family (Candidatus Brockarchaeota)
RLI45059.1 (Candidatus Bathyarchaeota)
RLF14881.1 (Thermoprotei archaeon)
NVM21655.1 CPBP family (Desulfobacterales bacterium)
RLI47487.1 (Candidatus Bathyarchaeota)
RLI00060.1 (Candidatus Bathyarchaeota)
RLI00222.1 (Candidatus Bathyarchaeota)
WP 008894482.1 (Haloterrigena salina)
WP 008424255.1 (Natronococcus jeotgali)
WP 008453115.1 (Natrinema)
WP 007738867.1 (Halococcus thailandensis)
WP 008389131.1 (Halogeometricum pallidum)
WP 007545129.1 (Haloferax larsenii])
WP 008323263.1 (Haloferax)
WP 008321658.1 (Haloferax mucosum)
WP 008577561.1 (unclassified Haloferax)
93
97
96
36
98
100
95
100
99
45
94
100
100
100
100
92
98
100
100
99
94
100
100
96
47
100
67
91
94
97
100
99
100
46
93
81
95
87
92
86
90
98
59
77
99
84
100
79
100
94
100
94
Tree scale: 1
Archaea
Eukaryotes
Bacteria
Atabeya-plasmid

Figure S6 Maximum likelihood phylogeny of CAAX proteases identified in the Atabeyarchaeia-2 plasmid_700. For this analysis, the top 50 protein hits from the NCBI database were selected. Sequence alignment was performed using MAFFT (version 7.310) with the 'auto' setting, and the alignment was subsequently trimmed with trimAl (version 1.4.rev15) using a gap threshold of 0.7 (-gt 0.7). The phylogenetic tree was initially constructed with IQ-TREE (version 1.6.1) employing the LG+FO+R model.

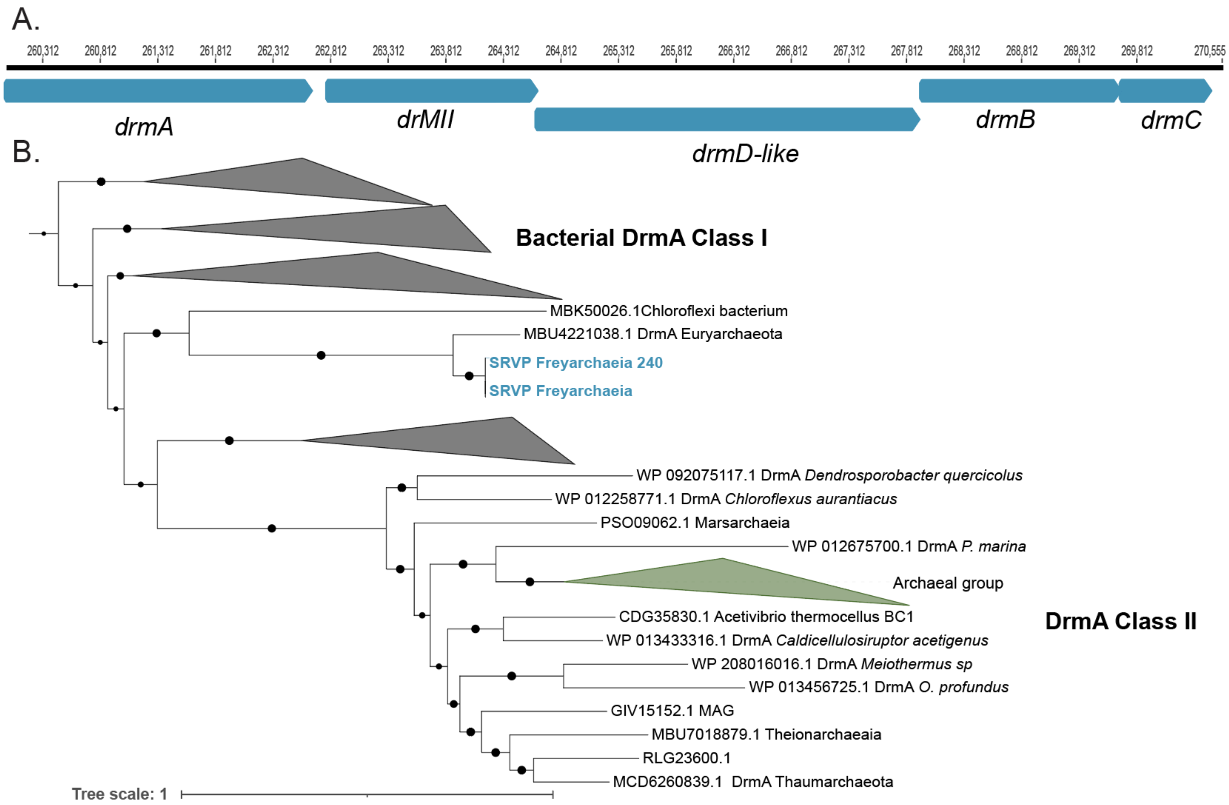

A.
260,312
260,812
261,312
261,812
262,312
262,812
263,312
263,812
264,312
264,812
265,312
265,812
266,312
266,812
267,312
267,812
268,312
268,812
269,312
269,812
270,555
drmA
drMII
drmD-like
drmB
drmC
B.
Bacterial DrmA Class I
MBK50026.1Chloroflexi bacterium
MBU4221038.1 DrmA Euryarchaeota
SRVP Freyarchaeia 240
SRVP Freyarchaeia
WP 092075117.1 DrmA *Dendrosporobacter quercicolus*
WP 012258771.1 DrmA *Chloroflexus aurantiacus*
PSO09062.1 Marsarchaeia
WP 012675700.1 DrmA *P. marina*
Archaeal group
DrmA Class II
CDG35830.1 Acetivibrio thermocellus BC1
WP 013433316.1 DrmA *Caldicellulosiruptor acetigenus*
WP 208016016.1 DrmA *Meiothermus sp*
WP 013456725.1 DrmA *O. profundus*
GIV15152.1 MAG
MBU7018879.1 Theionarchaeaia
RLG23600.1
MCD6260839.1 DrmA Thaumarchaeota
Tree scale: 1

Figure S7 Genomic organization of DISARM (Defense Island System Associated with Restriction-Modification) system from Freyarchaeia and phylogenetic placement of the A DrmA within the whole superfamily. A) DrmA, DrmMII, DrmA', DrmB and DrmC gene locus. B) Phylogenetic analysis of the DrmA protein found within the Freyarchaeia genome and reference sequences. The phylogenetic tree was initially constructed with IQ-TREE (version 1.6.1) employing the LG+FO+R model.

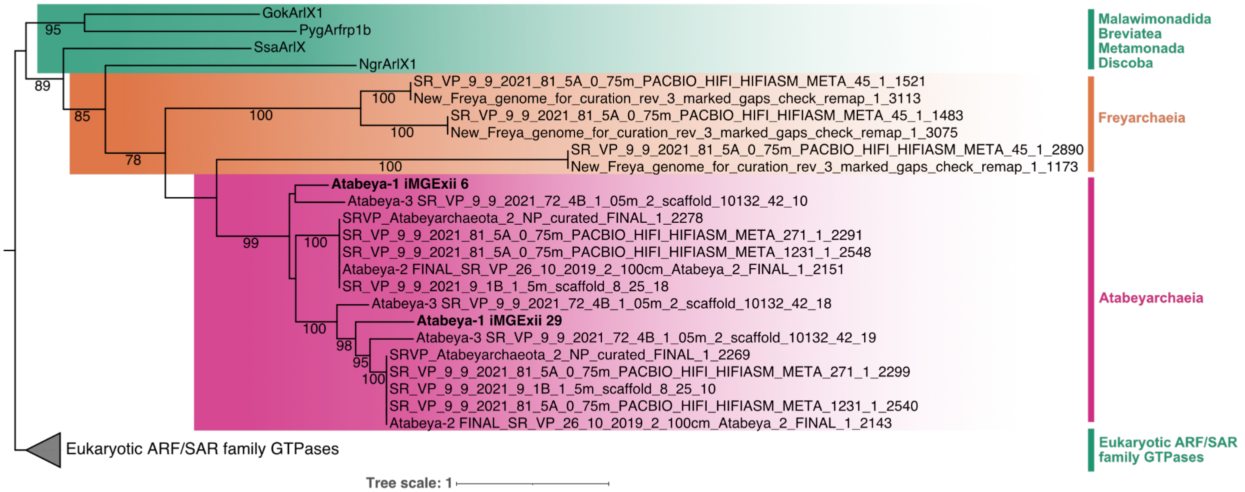

GokArlX1
PygArfrp1b
SsaArlX
NgrArlX1
95
89
85
78
100
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_45_1_1521
100
New_Freya_genome_for_curation_rev_3_marked_gaps_check_remap_1_3113
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_45_1_1483
100
New_Freya_genome_for_curation_rev_3_marked_gaps_check_remap_1_3075
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_45_1_2890
100
New_Freya_genome_for_curation_rev_3_marked_gaps_check_remap_1_1173
Atabeya-1 iMGExii 6
Atabeya-3 SR_VP_9_9_2021_72_4B_1_05m_2_scaffold_10132_42_10
SRVP_Atabeyarchaeota_2_NP_curated_FINAL_1_2278
99
100
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_271_1_2291
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_1231_1_2548
Atabeya-2 FINAL_SR_VP_26_10_2019_2_100cm_Atabeya_2_FINAL_1_2151
SR_VP_9_9_2021_9_1B_1_5m_scaffold_8_25_18
Atabeya-3 SR_VP_9_9_2021_72_4B_1_05m_2_scaffold_10132_42_18
100
Atabeya-1 iMGExii 29
98
Atabeya-3 SR_VP_9_9_2021_72_4B_1_05m_2_scaffold_10132_42_19
95
SRVP_Atabeyarchaeota_2_NP_curated_FINAL_1_2269
100
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_271_1_2299
SR_VP_9_9_2021_9_1B_1_5m_scaffold_8_25_10
SR_VP_9_9_2021_81_5A_0_75m_PACBIO_HIFI_HIFIASM_META_1231_1_2540
Atabeya-2 FINAL_SR_VP_26_10_2019_2_100cm_Atabeya_2_FINAL_1_2143
Eukaryotic ARF/SAR family GTPases
Tree scale: 1
Malawimonadida
Breviatea
Metamonada
Discoba
Freyarchaeia
Atabeyarchaeia
Eukaryotic ARF/SAR family GTPases

Figure S8 Comparison of GTPases encoded within Atabeyarchaeia putative MGEs and those integrated into the genomes of other Asgard archaea and Eukaryotes. Maximum likelihood tree of eukaryotic ARF family GTPases and Asgard archaea GTPase sequences. GTPases on the putative iMGE-xii are highlighted in bold. A curated set of proteins for the eukaryotic Arf family described in Vargová et al., (2021) was used to search in both Atabeyarchaeia and Freyarchaeia genomes and MGEs. A non-redundant subset of the references and Asgard hits with greater than 25% protein identity were aligned and trimmed with Mafft auto (v7.310) and trimAl-gt 0.5 (v1.4.rev15). The final tree was produced with iqtree (v.1.6.1) and LG+R9 model was chosen according to BIC.

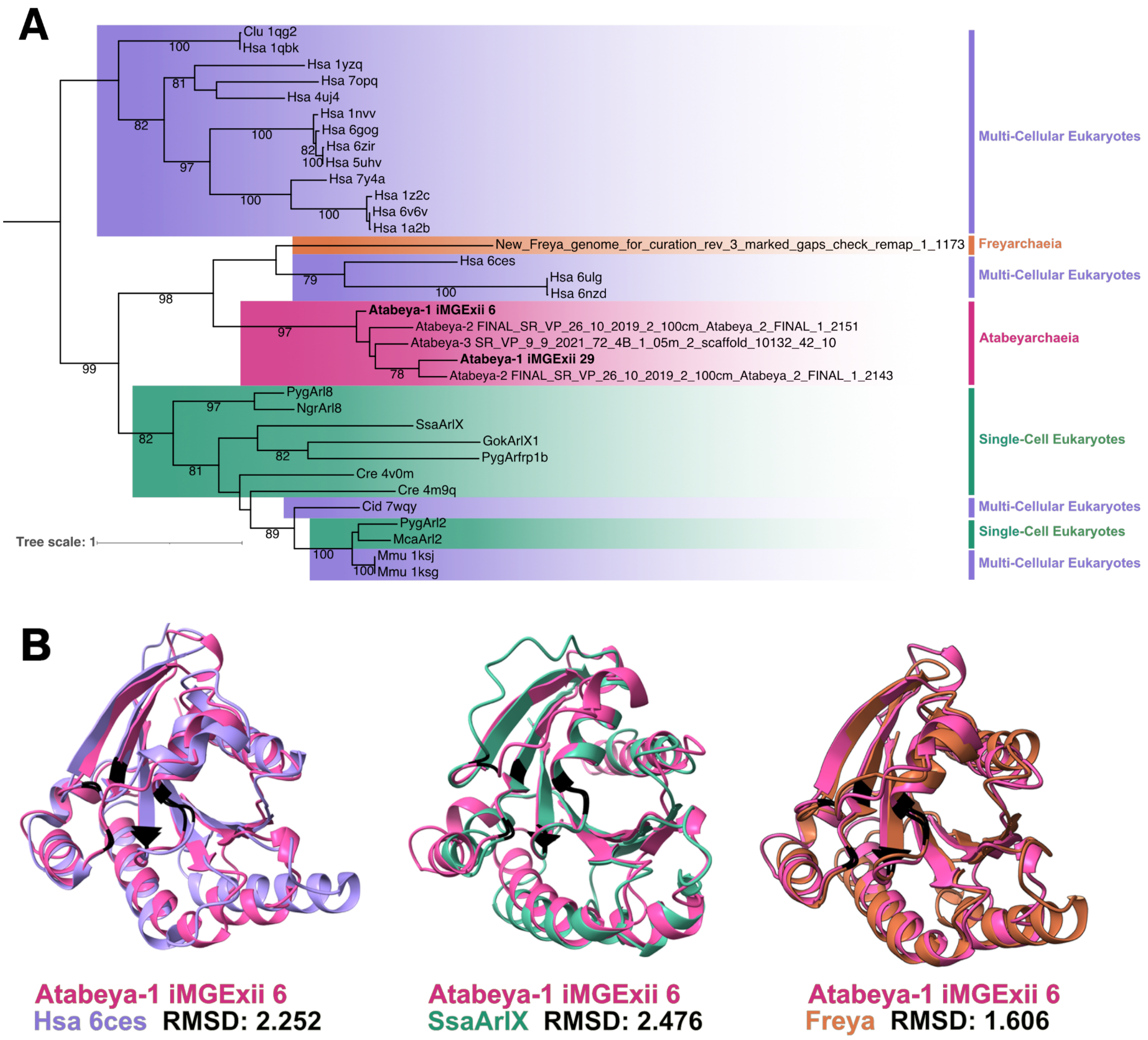


Figure S9 Structural similarity between Atabeyarchaeia MGE and eukaryotic GTPases. A) Structural phylogeny including a subset of the proteins with the sequence phylogeny and confirmed structural homologues in the RCSB Protein Data Bank (See "Methods"). The first three letters of the references in sequence and structural phylogenies are the first letter of the genus followed by the first two letters of the species (i.e., Hsa abbreviated for *Homo sapien*). B) Atabeya-1 iMGE-xii_6 GTPase structural model (pink) superimposed on a *H. sapien* GTPase predicted by electron microscopy (purple), *Spironucleus salmonicida* GTPase structural model (green), and Freyarchaeia GTPase structural model (orange) (from left to right).

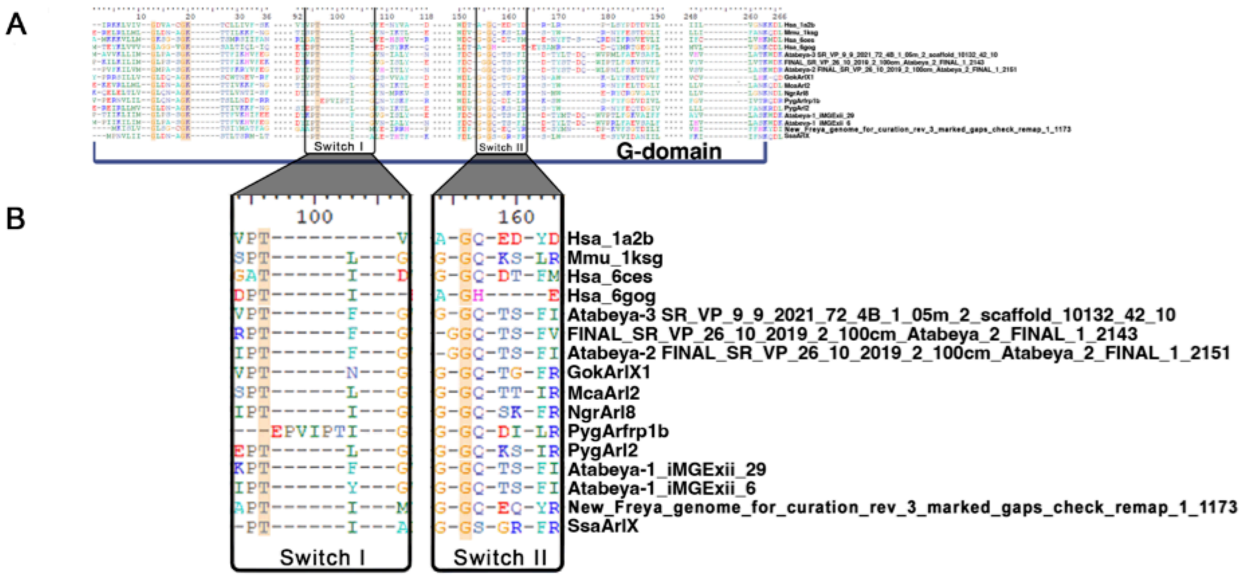
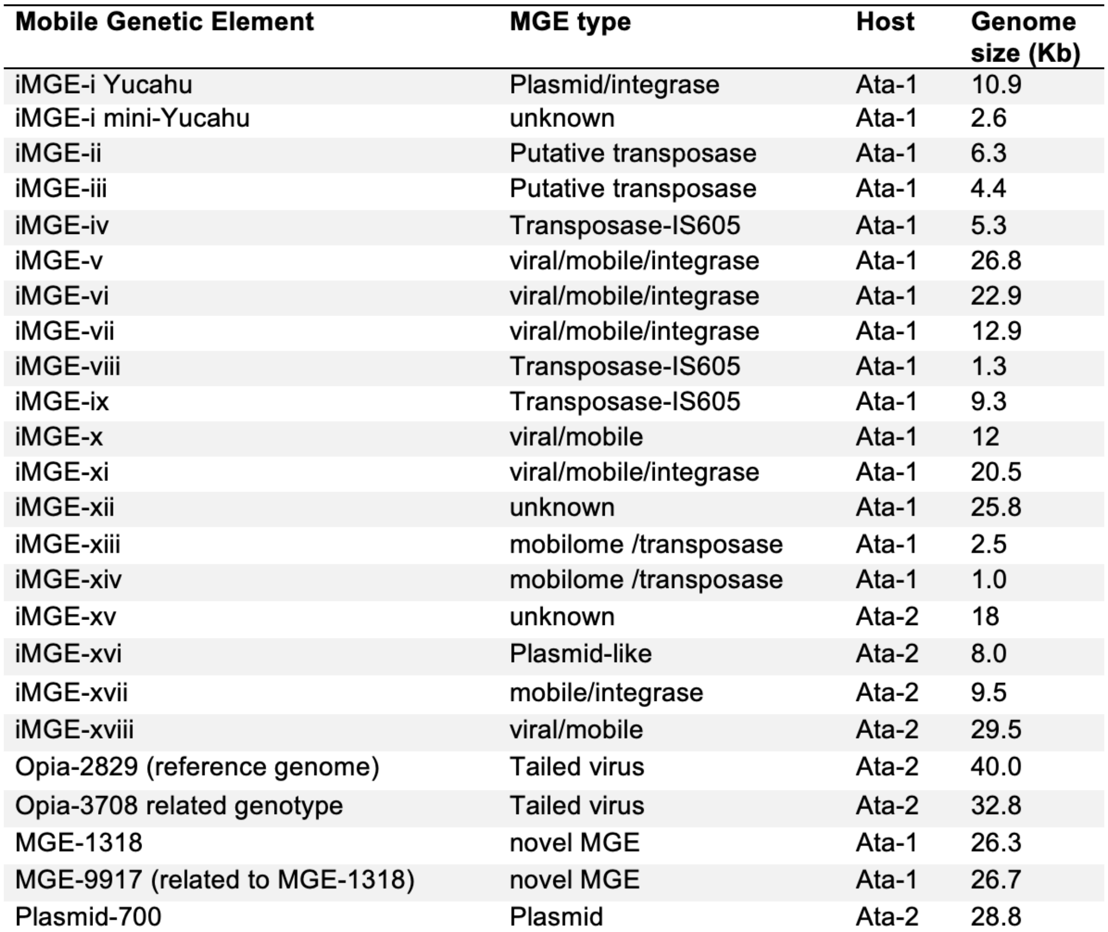

A
Switch I
Switch II
G-domain
B
100
160
VPT----------V A-GQ-ED-YD Hsa_1a2b
SPT------L---G G-GQ-KS-LR Mmu_1ksg
GAT------I---D G-GQ-DT-FM Hsa_6ces
DPT------I---- A-GH-----E Hsa_6gog
VPT------F---G G-GQ-TS-FI Atabeya-3 SR_VP_9_9_2021_72_4B_1_05m_2_scaffold_10132_42_10
RPT------F---G -GGQ-TS-FV FINAL_SR_VP_26_10_2019_2_100cm_Atabeya_2_FINAL_1_2143
IPT------F---G -GGQ-TS-FI Atabeya-2 FINAL_SR_VP_26_10_2019_2_100cm_Atabeya_2_FINAL_1_2151
VPT------N---G G-GQ-TG-FR GokArlX1
SPT------L---G G-GQ-TT-IR McaArl2
IPT------F---G G-GQ-SK-FR NgrArl8
---EPVIPTI---G G-GQ-DI-LR PygArfrp1b
EPT------L---G G-GQ-KS-IR PygArl2
KPT------F---G G-GQ-TS-FI Atabeya-1_iMGExii_29
IPT------Y---G G-GQ-TS-FI Atabeya-1_iMGExii_6
APT------I---M G-GQ-EQ-YR New_Freya_genome_for_curation_rev_3_marked_gaps_check_remap_1_1173
-PT------I---A G-GS-GR-FR SsaArlX
Switch I
Switch II

Figure S10 Conservation of G domain and switches. A) Multi-sequence structural alignment, highlighting the conserved amino acids (yellow) in the G domain and B) the position of the switches.

4.7 Tables



Mobile Genetic Element	MGE type	Host	Genome size (Kb)
iMGE-i Yucahu	Plasmid/integrase	Ata-1	10.9
iMGE-i mini-Yucahu	unknown	Ata-1	2.6
iMGE-ii	Putative transposase	Ata-1	6.3
iMGE-iii	Putative transposase	Ata-1	4.4
iMGE-iv	Transposase-IS605	Ata-1	5.3
iMGE-v	viral/mobile/integrase	Ata-1	26.8
iMGE-vi	viral/mobile/integrase	Ata-1	22.9
iMGE-vii	viral/mobile/integrase	Ata-1	12.9
iMGE-viii	Transposase-IS605	Ata-1	1.3
iMGE-ix	Transposase-IS605	Ata-1	9.3
iMGE-x	viral/mobile	Ata-1	12
iMGE-xi	viral/mobile/integrase	Ata-1	20.5
iMGE-xii	unknown	Ata-1	25.8
iMGE-xiii	mobilome /transposase	Ata-1	2.5
iMGE-xiv	mobilome /transposase	Ata-1	1.0
iMGE-xv	unknown	Ata-2	18
iMGE-xvi	Plasmid-like	Ata-2	8.0
iMGE-xvii	mobile/integrase	Ata-2	9.5
iMGE-xviii	viral/mobile	Ata-2	29.5
Opia-2829 (reference genome)	Tailed virus	Ata-2	40.0
Opia-3708 related genotype	Tailed virus	Ata-2	32.8
MGE-1318	novel MGE	Ata-1	26.3
MGE-9917 (related to MGE-1318)	novel MGE	Ata-1	26.7
Plasmid-700	Plasmid	Ata-2	28.8

Table 1 Manually curated genomes for chromosomally integrated genetic elements and extrachromosomal elements. MGE type is the closest classification based on gene content and genome structure. The host was determined based on matches of the MGE sequences to spacers from the CRISPR systems.

Sample	motifString	modificationType
Atabeyarchaeia 1	GNNGANNNNNNNRTTC	m6A
Atabeyarchaeia 1	CCGG	m4C
Atabeyarchaeia 1	GGCC	m4C
Atabeyarchaeia 1	CGCG	m4C
Atabeyarchaeia 1	GTSAC	m6A
Atabeyarchaeia 1	CTAG	m4C
Atabeyarchaeia 1	CTSAG	m6A
Atabeyarchaeia 1	CCAGG	m6A
Atabeyarchaeia 1	DGCGCH	m4C
Atabeyarchaeia 1	GAGGAA	m6A
Atabeyarchaeia 1	GATC	m6A
Atabeyarchaeia 1	GTAC	m4C
Atabeyarchaeia 1	GYATGAG	m6A
Atabeyarchaeia 2	GTSAC	m6A
Atabeyarchaeia 2	CAGNNNNNRTGG	m6A
Atabeyarchaeia 2	CGCG	m4C
Atabeyarchaeia 2	GATC	m6A
Atabeyarchaeia 2	CTAG	m4C
Atabeyarchaeia 2	ACAGG	m6A
Atabeyarchaeia 2	AGSCT	m4C
Atabeyarchaeia 2	DGCGCH	m4C
Atabeyarchaeia 2	GGCAG	m6A
Atabeyarchaeia 2	GRATGAG	m6A
Atabeyarchaeia 2	GTAC	m4C
Freyarchaeia strain	CCGG	m4C
Freyarchaeia strain	GGCC	m4C
Freyarchaeia strain	CCWGG	m4C
Freyarchaeia strain	GCGC	m4C
Freyarchaeia strain	GGWCC	m4C
Freyarchaeia strain	CCGG	m4C

Table 2 Methylation patterns in Atabeyarchaeia and Freyarchaeia genomes based on PacBio data.

4.8 References

Al-Shayeb, B., Schoelmerich, M. C., West-Roberts, J., Valentin-Alvarado, L. E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M. J., Williams, K. H., Doudna, J. A., & Banfield, J. F. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature*, *610*(7933), 731–736.

Anton, B. P., & Roberts, R. J. (2021). Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes. *Annual Review of Microbiology*, *75*, 129–149.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., Froula, J., Kang, D. D., Malmstrom, R. R., Morgan, R. D., Posfai, J., Singh, K., Visel, A., Wetmore, K., Zhao, Z., Rubin, E. M., Korlach, J., Pennacchio, L. A., & Roberts, R. J. (2016). The Epigenomic Landscape of Prokaryotes. *PLoS Genetics*, *12*(2), e1005854.

Bravo, J. P. K., Aparicio-Maldonado, C., Nobrega, F. L., Brouns, S. J. J., & Taylor, D. W. (2022). Structural basis for broad anti-phage immunity by DISARM. *Nature Communications*, *13*(1), 2987.

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/servlets/purl/1241166

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315–333.

Dong, R., Peng, Z., Zhang, Y., & Yang, J. (2018). mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* , *34*(10), 1719–1725.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., & Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, *359*(6379). https://doi.org/10.1126/science.aar4120

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), 992–999.

Feng, X., Cheng, H., Portik, D., & Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, *19*(6), 671–674.

Ghaly, T. M., Tetu, S. G., Penesyan, A., Qi, Q., Rajabal, V., & Gillings, M. R. (2022). Discovery of integrons in Archaea: Platforms for cross-domain gene transfer. *Science Advances*, *8*(46), eabq6376.

Gomis-Rüth, F. X., Solá, M., Acebo, P., Párraga, A., Guasch, A., Eritja, R., González, A., Espinosa, M., del Solar, G., & Coll, M. (1998). The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. *The EMBO Journal*, *17*(24), 7404–7415.

Guo, X., & Huang, L. (2010). A superfamily 3 DNA helicase encoded by plasmid pSSVi from the hyperthermophilic archaeon Sulfolobus solfataricus unwinds DNA as a higher-order oligomer and interacts with host primase. *Journal of Bacteriology*, *192*(7), 1853–1864.

Hug, L. A., Castelle, C. J., Wrighton, K. C., Thomas, B. C., Sharon, I., Frischkorn, K. R., Williams, K. H., Tringe, S. G., & Banfield, J. F. (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome*, *1*(1), 22.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kieft, K., & Anantharaman, K. (2022). Deciphering Active Prophages from Metagenomes. *mSystems*, *7*(2), e0008422.

Krupovic, M., Makarova, K. S., Wolf, Y. I., Medvedeva, S., Prangishvili, D., Forterre, P., & Koonin, E. V. (2019). Integrated mobile genetic elements in Thaumarchaeota. *Environmental Microbiology*, *21*(6), 2056–2078.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

Medvedeva, S., Sun, J., Yutin, N., Koonin, E. V., Nunoura, T., Rinke, C., & Krupovic, M. (2022). Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. *Nature Microbiology*, *7*(7), 962–973.

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H., & Ferrin, T. E. (2023). UCSF ChimeraX: Tools for structure building and analysis. *Protein Science: A Publication of the Protein Society*, *32*(11), e4792.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682.

Mizuno, C. M., Prajapati, B., Lucas-Staat, S., Sime-Ngando, T., Forterre, P., Bamford, D. H., Prangishvili, D., Krupovic, M., & Oksanen, H. M. (2019). Novel haloarchaeal viruses from Lake Retba infecting Haloferax and Halorubrum species. *Environmental Microbiology*, *21*(6), 2129–2147.

Moreno-Cinos, C., Goossens, K., Salado, I. G., Van Der Veken, P., De Winter, H., & Augustyns, K. (2019). ClpP Protease, a Promising Antimicrobial Target. *International Journal of Molecular Sciences*, *20*(9). https://doi.org/10.3390/ijms20092232

Morgan, R. D., Dwinell, E. A., Bhatia, T. K., Lang, E. M., & Luyten, Y. A. (2009). The MmeI family: type II restriction-modification enzymes that employ single-strand modification for host protection. *Nucleic Acids Research*, *37*(15), 5208–5221.

Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., Doron, S., & Sorek, R. (2018). DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nature Microbiology*, *3*(1), 90–98.

Payne, L. J., Meaden, S., Mestre, M. R., Palmer, C., Toro, N., Fineran, P. C., & Jackson, S. A. (2022). PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Research*, *50*(W1), W541–W550.

Rambo, I. M., Langwig, M. V., Leão, P., De Anda, V., & Baker, B. J. (2022). Genomes of six viruses that infect Asgard archaea from deep-sea sediments. *Nature Microbiology*, *7*(7), 953–961.

Raymann, K., Forterre, P., Brochier-Armanet, C., & Gribaldo, S. (2014). Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea.

Genome Biology and Evolution, *6*(1), 192–212.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2015). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, *43*(Database issue), D298–D299.

Rocha, E. P. C., & Bikard, D. (2022). Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biology*, *20*(1), e3001514.

Rubin, B. E., Diamond, S., Cress, B. F., Crits-Christoph, A., Lou, Y. C., Borges, A. L., Shivram, H., He, C., Xu, M., Zhou, Z., Smith, S. J., Rovinsky, R., Smock, D. C. J., Tang, K., Owens, T. K., Krishnappa, N., Sachdeva, R., Barrangou, R., Deutschbauer, A. M., … Doudna, J. A. (2022). Species- and site-specific genome editing in complex bacterial communities. *Nature Microbiology*, *7*(1), 34–47.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *The CRISPR Journal*, *3*(6), 462–469.

Seitz, K. W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J. R., Teske, A. P., Ettema, T. J. G., & Baker, B. J. (2019). Asgard archaea capable of anaerobic hydrocarbon cycling. *Nature Communications*, *10*(1), 1822.

Senčilo, A., & Roine, E. (2014). A Glimpse of the genomic diversity of haloarchaeal tailed viruses. *Frontiers in Microbiology*, *5*, 84.

Shen, B. W., Heiter, D. F., Chan, S.-H., Wang, H., Xu, S.-Y., Morgan, R. D., Wilson, G. G., & Stoddard, B. L. (2010). Unusual target site disruption by the rare-cutting HNH restriction endonuclease PacI. *Structure* , *18*(6), 734–743.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179.

Speth, D. R., Yu, F. B., Connon, S. A., Lim, S., Magyar, J. S., Peña-Salinas, M. E., Quake, S. R., & Orphan, V. J. (2022). Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *The ISME Journal*, *16*(7), 1750–1764.

Takai, Y., Sasaki, T., & Matozaki, T. (2001). Small GTP-binding proteins. *Physiological Reviews*, *81*(1), 153–208.

Tamarit, D., Caceres, E. F., Krupovic, M., Nijland, R., Eme, L., Robinson, N. P., & Ettema, T. J. G. (2022). A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nature Microbiology*, *7*(7), 948–952.

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., & Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, *13*(1), 2561.

Valentin-Alvarado, L. E., Appler, K. E., De Anda, V., Schoelmerich, M. C., West-Roberts, J., Kivenson, V., Crits-Christoph, A., Ly, L., Sachdeva, R., Savage, D. F., Baker, B. J., & Banfield, J. F. (2023). Asgard archaea modulate potential methanogenesis substrates in wetland soil. In *bioRxiv* (p. 2023.11.21.568159). https://doi.org/10.1101/2023.11.21.568159

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., & Steinegger, M. (2023). Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01773-0

Vetter, I. R., & Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science*, *294*(5545), 1299–1304.

Wein, T., & Sorek, R. (2022). Bacterial origins of human cell-autonomous innate immune mechanisms. *Nature Reviews. Immunology*, *22*(10), 629–638.

Wu, F., Speth, D. R., Philosof, A., Crémière, A., Narayanan, A., Barco, R. A., Connon, S. A., Amend, J. P., Antoshechkin, I. A., & Orphan, V. J. (2022). Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularized Asgard archaea genomes. *Nature Microbiology*, *7*(2), 200–212.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., & Ettema, T. J. G. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, *541*(7637), 353–358.

Transitional section

The analysis of mobile genetic elements (MGEs) associated with Asgard archaeal genomes in Chapter 4 highlighted the genetic versatility that may have facilitated the emergence of eukaryotic features from an archaeal ancestor. However, gaining a comprehensive understanding of the diversity and evolution of archaea and their roles in complex environmental microbiomes requires recovering complete genomes from metagenomic data. In Chapter 5, I leveraged the latest long-read sequencing (LRS) technologies, specifically the PacBio Revio system, to sequence wetland soil samples and assemble hundreds of circularized, complete, or near-complete genomes from archaea, bacteria, and their associated genetic elements. This transition from relying on short-read Illumina sequencing to now utilizing long-read PacBio sequencing represents a new era in metagenomics, enabling the recovery of high-quality genomes and associated mobile elements from complex environmental samples. The availability of these complete genomes facilitated precise measurements of genome architecture, gene content, and the mapping of transcriptomic data to link gene expression patterns to specific organisms within the microbiome. This powerful combination of long-read metagenomics and metatranscriptomics, as described in Chapter 5, provides insights into the structure, function, and activity of diverse microbial communities, including archaea playing crucial roles in wetland biogeochemical cycles.

5. Wetland soil hosts a high diversity of novel archaea actively expressing pathways for carbon cycling and represented by complete genomes

Luis E. Valentin-Alvarado, Daniel M. Portik, Jeremy E. Wilkinson, David F. Savage and Jillian F. Banfield.

Unpublished, 2023.

Abstract

Long-read sequencing (LRS) technologies, particularly PacBio HiFi reads, have significantly advanced the field of metagenomics by enabling the assembly of more complete genomes. The new PacBio Revio system generates very high fidelity (HiFi) reads (>Q20), necessitating the development of new assembly algorithms, such as hifiasm-meta and metaMDBG. Here, we applied Revio HiFi sequencing and these assembly methods to investigate wetland soil microbiomes and generated >1,300 metagenome-assembled genomes (MAGs), including more than 400 that are circularized and either complete (if validated using Illumina reads) or potentially complete. Included are 1,000 bacterial genomes (most prevalent are those from Acidobacteria) and 150 genomes from diverse archaea that dominate in deep soil. Interestingly, metaMDBG recovered over 3,000 circular contigs that are less than 200 kb. We focused our analyses on genomes from 110 archaeal groups: Bathyarchaeia, Methanosarsinales, Thermoplasmatales, Thaumarchaeota, Hadearchaeales, Methanomethyliales, methanogens, Asgardarchaeota and DPANN archaea. The availability of complete genomes enabled exact measures of genome size, genome architecture, gene content and precise linkage between RNA transcripts and the organisms from which they originate. Metatranscript mapping revealed highly expressed MCR genes from Methanosarsinales, Methanomicrobiales, and Methanomethyliales. These findings highlight the power of long-read metagenomics to recover complete microbial genomes from complex environments, enabling insights into microbial community structure, function, and gene expression patterns.

***N.B.** All main figures and tables for this manuscript can be found below in sections 5.5 and 5.6.*

5.1 Introduction

Archaea are now recognized as common members of many ecosystems including sediments, the deep ocean, animal microbiomes, hot springs and soil, although sometimes their relative abundance levels are low. The metabolic capacities of archaea vary substantially, and include the ability to produce or consume methane, generate energy by degradation of proteins or complex hydrocarbons, or oxidize or reduce sulfur compounds (Blohs et al. 2019). Often, their activities interconnect multiple biogeochemical cycles. Despite their ecological importance, the full scope of archaeal diversity and function remains poorly understood, partly due to the historical bias towards culturing and genomically studying methanogens and archaea that inhabit extreme environments. Particularly underrepresented are genomes of uncultivated soil-associated archaea, thus understanding of their metabolic capabilities and ecosystem roles is quite limited.

Wetland soils are important carbon sinks and sources of greenhouse gases, yet the roles of archaea in these processes remain poorly understood. Wetlands store approximately 20-30% of the global soil carbon and contribute significantly to methane emissions (Kayranli et al. 2010; Guerra et al. 2020). Archaea, particularly methanogens, are known to play crucial roles in carbon cycling and methane production in wetland soils (Evans et al. 2019; Angle et al. 2017; Oliverio et al. 2024). However, the diversity and metabolic capabilities of wetland soil archaea beyond methanogens remain largely unexplored.

The field of genome-resolved metagenomics has provided a powerful lens through which to view the microbial world, bypassing the need for cultivation and allowing for the direct analysis of genetic material from environmental samples (Tyson et al. 2004). This has led to the discovery of a vast array of previously unknown archaeal lineages, expanding our view of their diversity and ecological roles (Rinke et al. 2013; Brown et al. 2015). However, the challenge of assembling high-quality, complete genomes from metagenomic data is considerable. This is particularly problematic for complex microbial communities, such as those from soil, and generally results in fragmented and incomplete genomes (Quince et al. 2017; Eisenhofer et al. 2023). Contamination of draft genome bins with genome fragments from other organisms restricts confidence in metabolic predictions. While the number of archaeal metagenome-assembled genomes (MAGs) has grown to approximately 20,000, less than 3% are considered complete. However, most of these are still fragmented, as the completeness metric generally relies on counts of expected genes not reconstruction of circularized, finished sequences. Very few are closed and validated throughout. Despite quality limitations, a great deal has been learned from partial archaeal MAGs, including the discovery of the importance of Asgard archaea in eukaryogenesis (Spang et al. 2015; Eme et al. 2023) and the prevalence of episymbiotic ultra-small archaea in many ecosystems (Castelle et al. 2018; Dombrowski et al. 2020; He et al. 2021).

The availability of complete archaeal genomes greatly improves our ability to decipher the metabolic pathways thus to predict ecological functions, to identify sets of integrated genetic elements (Chapter 4), to document genome size and structure and to reconstruct histories. Here, we leveraged new, highly accurate long-read sequencing to reconstruct over a thousand MAGs from deep (>60 cm) wetland soil. To enable in-depth analyses, we focused our analyses on only a subset of 118 of the >500 circularized genomes that are from archaea, which we find are highly prevalent in deep anaerobic soil regions. To date, few studies have extensively deployed accurate long-read sequencing to study complex microbiomes. This study contrasts with a prior study, in which the previous (less accurate) type of PacBio Sequel sequencing was used to recover 428 high-quality MAGs, of which only 28 were potentially circular (Bickhart et al. 2022). Unlike the current study that targeted soil, these 428 genomes were reconstructed from a relatively low-complexity fecal sample of an adult sheep and relied on Hi-C data for binning. The complete (validated and, if necessary, polished using Illumina short read-based curation) and effectively complete (circularized but not completely validated) genomes generated in the current study are among the first such high-quality genomes for several major archaeal groups. We used these genomes to provide context for metatranscriptomic analyses used to identify highly expressed genes in each archaeon, providing insights into the roles of specific archaea in the deep, wetland soil ecosystem.

Our work provides (1) a genomic catalog of complete archaeal genomes from wetland soils, overcoming limitations of short-read-based draft genomes; (2) an expanded understanding of archaeal diversity and function in these ecosystems, complementing the bacterial focus of previous studies; (3) detailed metabolic analyses supported by metatranscriptomics, providing accurate context for in situ activity; and (4) insights into soil microbial diversity in the context of climate change. This study demonstrates the power of combining accurate long-read sequencing with metatranscriptomics to unravel the complexity of wetland soil microbiomes and highlights the ecological importance of previously understudied archaeal lineages.

5.2 Materials and Methods

Long-read DNA extraction and sequencing and assembly

Six soil samples were collected on November 14th, 2023, from depths of 60, 70, 80, 90, 100, and 110 cm. The samples were flash-frozen using liquid nitrogen and stored at -80°C until processing. For general information about the sampling protocol, refer to the methods sections of previously published work (Al-Shayeb et al. 2022; Valentin-Alvarado et al. 2023).The samples were extracted using a modified version of the PowerSoil DNA Extraction Kit (Qiagen). Briefly, the samples were thawed, and 5 mL of BPER-II was added per 4 g of soil. The samples were then mixed with the beads from the PowerSoil kit and placed in a water bath at 65°C for 30 minutes, with gentle inversion every 5 minutes. This adjustment was necessary to avoid the vortexing step that could potentially shear the DNA. After this step, the subsequent steps from the PowerSoil kit were followed. Genomic DNA was extracted from 6 samples collected at different depths (60cm, 80cm, 90cm, 100cm, 110cm, 115cm). 3 µg of each DNA sample was sheared to a target size of >3 kb using a Megaruptor 3 instrument (Diagenode) at a speed setting of 29. Libraries were prepared using the SMRTbell prep kit 3.0 (PacBio) following the manufacturer's protocol. This included ligation of barcoded adapters for multiplexing. Size selection was performed using 3.1X 35% AMPure beads to deplete DNA fragments <3 kb. Good library yields were obtained for all 6 samples. Libraries were pooled in equal mass to generate 3 pools each containing 2 samples with similar size distributions (Pool 1: 60cm, 90cm; Pool 2: 80cm, 100cm; Pool 3: 110cm, 115cm). Each pooled library was loaded at a concentration of 175 pM on a PacBio Revio system and sequenced using a 24-hour movie collection time. This generated >2.5 million high-fidelity (HiFi) reads per sample. Reads were quality trimmed using BBDuk (bbduk.sh minavgquality=20 qtrim=rl trimq=20) (Bushnell 2014) and assembled with hifiasm-meta (Feng et al. 2022) and metaMDBG (Benoit et al. 2024).

Taxonomic composition classification and relative abundance

To determine the taxonomic composition and estimate the relative abundances, the following approach was employed: The dereplicated high-quality PacBio archaeal genomes were combined with other high-quality metagenome-assembled genomes (MAGs) obtained from the Illumina dataset. All high-quality genomes present in the vernal pool samples were mapped independently to all the metagenome reads using BBMap (Bushnell 2014). The mapping parameters were set as follows: nodisk=t, pigz=t, unpigz=t, ambiguous=random. The minimum identity threshold for mapping was set to 0.9, and ambiguous mappings were handled randomly. The relative abundance of each genome, as well as the percentage of unmapped reads, was calculated using coverM1 with the following parameters: -m relative_abundance --min-read-percent-identity 95. This approach allowed for the taxonomic classification of the genomes present in the samples and the estimation of their relative abundances based on the mapping of metagenome reads to the high-quality reference genomes. The use of both PacBio and Illumina datasets ensured a comprehensive representation of the microbial community, while the mapping parameters and tools were carefully chosen to ensure accurate and reliable results.

Metabolic reconstruction of high quality circular and potentially complete archeal genomes

We utilized Prodigal (v2.6.3) to predict the genes for 118 high-quality archaeal genomes. A suite of databases, including KofamKOALA (v1.3.0) and InterProScan (5.50-84.0) were combined to begin annotation. Complete genomes were also annotated with HydDB (Søndergaard et al. 2016), METABOLIC (v4.0) (Zhou et al. 2022), PROKKA (v1.14.6), DRAM (Shaffer et al. 2020), and MEBS (v2.0).

Phylogenetic analysis of 15 concatenated ribosomal proteins of high-quality archaeal genomes

Concatenated amino acid sequences of 15 ribosomal proteins (RP15) were extracted from high-quality metagenome-assembled genomes (MAGs) reconstructed from soil samples collected at the Sevilleta National Wildlife Refuge (New Mexico, USA). These RP15 sequences were combined with reference archaeal genome sequences obtained from the Genome Taxonomy Database (GTDB). Multiple sequence alignments were generated and used to infer a maximum likelihood phylogenetic tree with IQ-TREE software (Minh et al. 2013). The best-fit evolutionary model LG+C60+F was automatically selected and ultrafast bootstrap support values were calculated to assess branch robustness. The resulting tree was visualized and annotated using the Interactive Tree of Life (iTOL) online tool.

5.3 Results and Discussion

Archaea dominate deep wetland soil

Analysis of 88 metagenomic datasets indicates that bacteria dominate the wetland topsoil layers, whereas Archaea become more prevalent than bacteria below a soil depth of 60 cm (Figure 5.1A). In the deepest soil zones, Thermoprotea archaea (Bathyarchaeia, Nitrospiria, Methanomethylicia and Brockarchaeia) account for around 50% of the entire microbial community. The increase is paralleled by a modest increase in abundance of Asgard archaea and a notable general increase in abundance of Chloroflexi bacteria (Figure 5.1B).

Wetland soil PacBio metagenome assembly produced hundreds of circular contigs

We applied HiFi Revio sequencing (2plex setup on the Revio SMRT Cell; see “methods”) to samples collected from wetland soil (the SRVP site in Lake County, northern California (Al-Shayeb et al. 2022; Valentin-Alvarado et al. 2023). The sequencing targeted 6 samples taken from soil depths of 60, 80, 90, 100, 110, and 115 cm, where the soil is permanently wet. The 20,661,987 HiFi reads generated corresponded to 195.88 Gigabases of DNA sequence information (Table S1). Hifiasm-meta and metaMDBG bioinformatic tools were used to assemble the datasets from each sample independently, yielding 1,232 MAGs, of which 563 are circular PacBio sequences for bacteria and archaea of > 500 Kbp in length that are potentially complete. MetaMDBG assembly yielded 3,477 circular contigs with lengths of <200 kb, compared to 444 reconstructed by Hifiasm-meta. This suggests the superior capacity of MetaMDBG to identify potential mobile genetic elements (MGEs). We generated 58 circularized contigs between 200 kbp and 1 Mbp in length from Hifiasm-meta and 44 from MetaMDBG assemblies (Figure S1A-B). These represent candidate MGEs and some CPR bacteria and DPANN archaea. Hifiasm-meta and metaMDBG generated 358 and 312 circular bacterial and archaeal sequences, respectively, of >1Mb in length (Figure S2; Figure S3). This indicates comparable performance of the two algorithms for the assembly of larger genomic structures. The combined set of circularized microbial genomes from the two assemblies was dereplicated at 95% nucleotide identity (approximately species level. Of the dereplicated set, 118 archaeal microbial genomes of > 1 Mbp in length. A plot of GC content vs. genome size revealed a range in GC content from 32% to 68% and genome sizes up to > 5 Mbp (Figure 5.2). These genomes were assigned to 16 initial taxonomic groups and used for subsequent detailed phylogenetic analyses.

Phylogenetic classification of very high-quality Illumina genomes

We augmented the PacBio genomes with three Illumina-based complete genomes or near-complete MAGs that we did not recover PacBio sequences for (Table S2). From each genome in the combined dataset, 15 ribosomal proteins were concatenated and used for phylogenetic analysis (Figure 5.3). This revealed circularized genomes from 22 major archaeal groups (red in Figure 5.3) and four additional clades represented by MAGs (pink in Figure 5.3). At the approximate phylum level, we report high-quality, circularized and finished genomes from 18 archaeal lineages (colored groups on the outer ring of Figure 5.3).

The availability of essentially complete and complete, circularized genomes from a large range of archaeal lineages mostly lacking isolated representatives provided the opportunity to evaluate replication modes across a span of archaea. We calculated the GC skew and cumulative GC skew patterns and found that the majority (60%) of the cumulative GC skew patterns display very low skew and/or the patterns were too noisy to predict a replication mode. However, 25% of patterns provided variably strong indicators of typical bacterial-like (Chen et al. 2020) bidirectional replication from a single origin on two replichores of sub-equal lengths (Figure S4-A). Most of the remaining 15% of patterns were suggestive of rolling circular replication mode (Figure S4-B).

Archaeal community metabolic capacities and in situ activity

We established metabolic predictions for 58 circularized and/or finished genomes (Table S3). To facilitate analysis of the transcriptomic datasets acquired from the same samples as the PacBio DNA sequences, we selected the highest quality genome from each of the 18 major archaeal taxonomic groups and focused on the 10 genomes of metabolically interesting archaea (i.e., excluding most DPANN archaea) that display high transcriptional activity (Table S4). We predicted the *in situ* functions of most importance for each group (based on highly transcribed genes) and analyzed these findings in the context of the overall predicted metabolism (Figure 5.4).

Bathyarchaeia

Within the soil archaeal community, Bathyarchaeia is the most abundant group, representing 73.78% of the total archaeal abundance. Despite this, their gene expression levels are relatively low compared to gene expression levels for other archaeal genomes. The only genome for which we detected transcripts is the complete Bathyarchaeia-44 genome, so we chose this as the Bathyarchaeia representative for analysis of *in situ* activity. Expressed genes encode archaeal chaperonins (K22447), which are essential for protein folding and stress response. Also relatively highly expressed are genes involved in carbohydrate transport and metabolism, including those for a glucose/mannose transport system substrate-binding protein (srvp_genome_44.fna_1829, K17315) and multiple sugar transport system ATP-binding proteins (srvp_genome_44.fna_1826, K10112). Genes encoding components of the ABC-2 type transport system were also identified, again suggesting the importance of nutrient acquisition and ion transport.

Generally present in Bathyarchaeia are genes to break down cellulose to cellodextrin and cellobiose and for the storage of glucose as starch/glycogen. These archaea are likely to be able to convert formate and formaldehyde to CO_2 via the tetrahydrofolate methyl branch of the Wood–Ljungdahl pathway and to acetyl-CoA. They may be capable of fixing CO_2 via acetyl-CoA synthetase and use a form III Rubisco to incorporate CO_2 to form G3P via the nucleotide salvage pathway. Although they have many steps in the acetate-methane pathway, they lack the final genes needed to form methane; thus, these wetland soil Bathyarchaeia are not methanogens. They appear to convert glycine and serine to ammonia and pyruvate via the reductive glycine pathway (rGlyP). Consistent with the transcript data indicating expression of transporters, Bathyarchaeia have amino acid import/export genes. They have genes for the polar amino acid transport system substrate-binding protein (K02030), along with components of the branched-chain amino acid transport system, including ATP-binding proteins (K01996, K01995), a permease protein (K01997), and a substrate-binding protein (K01999). Additionally, two other genomes encode for a polar amino acid transport system ATP-binding protein (K02028) and a permease protein (K02029), indicating mechanisms for the uptake of both polar and branched-chain amino acids.

All wetland Bathyarchaeia have genomes that encode the acetate---CoA ligase (ADP-forming) subunits alpha (acdA K01905) and beta (acdB K22224), indicating the capacity to activate acetate into acetyl-CoA, a crucial step in both acetate formation and energy conservation. They also encode genes for the formation of acetate from acetyl-CoA via acetyl phosphate (K00625 and K01512) and directly via acetyl-CoA synthetase (K01895). They also have genes to interconvert alcohols and aldehydes (e.g., aldehyde:ferredoxin oxidoreductase; K03738). Only one

Bathyarchaeia has a propanol-preferring alcohol dehydrogenase (K13953). These two genes indicate the capacity for alcohol fermentation.

The Bathyarchaeia genomes have genes involved in sulfur metabolism, including sulfite reductase (K00392). The reduction of sulfite to sulfide is a key step in sulfur assimilation (and dissimilatory sulfate reduction). The presence of adenylylsulfate kinase (K00860) suggests the capability for sulfate activation, a precursor step in the biosynthesis of sulfur-containing compounds or in sulfate reduction processes. Cysteine synthase (K01738) indicates the synthesis of cysteine from sulfide. Additionally, sulfhydrogenase genes (K17996) may point to the reduction of elemental sulfur to hydrogen sulfide.

The Bathyarchaeia genomes encode a suite of enzymes involved in the beta-oxidation pathway. This pathway is for the catabolism of fatty acids into acetyl-CoA. Key enzymes identified include acetyl-CoA C-acetyltransferase (K00626) for the initial activation of acetyl units, enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase (K15016) and enoyl-CoA hydratase (K01715) for the hydration of enoyl-CoA intermediates, 3-hydroxybutyryl-CoA dehydrogenase (K00074) for the oxidation of hydroxyacyl-CoA, butyryl-CoA dehydrogenase (K00248), and acyl-CoA dehydrogenase (K00249) for the dehydrogenation of acyl-CoA to enoyl-CoA. Furthermore, the presence of the electron transfer flavoprotein subunits alpha (K03522) and beta (K03521) in at least one genome suggests a mechanism for transferring electrons generated during beta-oxidation to the and electron accepting complex, likely an ion-pumping hydrogenase. In terms of cycling of redox carriers, they may use group 3b NiFe-hydrogenases to oxidize NADPH to form H_2 (instead of H_2S) or alternatively 3c-hydrogen-dependent electron bifurcating system MvhADG-HdrABC.

A distinctive feature of these Bathyarchaeia genomes is the general absence of the prokaryotic ATPase synthase complex. Only two genomes encode complete V/A-type H+/Na+-transporting ATPase (K02117 through K02124). They also lack most components of an electron transport chain and likely form ATP by substrate-level phosphorylation.

Some Bathyarchaeia genomes encode enzymes involved in methylated compound metabolism, such as trimethylamine---corrinoid protein Co-methyltransferase (K14083), trimethylamine corrinoid protein (K14084), methylamine---corrinoid protein Co-methyltransferase (K16176), and dimethylamine corrinoid protein (K16179). Utilization of methylated nitrogen compounds may contribute to energy conservation and provide carbon and nitrogen sources.

Overall, the metabolism of soil Bathyarchaeia appears to center on the uptake and metabolism of a variety of sugars, fatty acids, methylated substrates and amino acids by producing acetate as a byproduct of their metabolism.

Thermoplasmata

The genome of a representative of the Thermoplasmatales (srvp_genome-48) has expressed genes for various cellular functions. The translation initiation factor 2 subunit 1 gene (K03237) is essential for the initiation of protein synthesis. The MoxR-like ATPase gene (K03924) may play a role in overcoming oxidative stress. The gene for 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase (K01737) is involved in synthesizing tetrahydrobiopterin, which is potentially linked to folate metabolism. Two genes for peptide/nickel transport system substrate-binding proteins (K02035) indicate active transport of peptides and nickel. Additionally, the nucleoside kinase gene (K22026) is implicated in nucleotide synthesis. The expressed FlaI gene (K07332) suggests that these archaea are motile in the soil environment.

Thermoplasmatales constitute 6.68% of the wetland soil archaeal community. We recovered 10 potentially complete genomes, representing 9 distinct lineages at the class/order level. Among the findings is the identification of a variant of the Wood–Ljungdahl pathway (WLP) for carbon fixation and energy conservation that is bacterial-like instead of the typical archaeal version. Two genomes (srvp_genome_50 and srvp_genome_81) encode the complete set of WLP enzymes, enabling the conversion of CO_2 to acetyl-CoA. The other genomes have only a subset of WLP components. They generally have genes encoding key enzymes in central carbon metabolism pathways that enable the consumption of pyruvate via the citric acid cycle (TCA cycle) and genes for β-oxidation of fatty acids.

We identified complete NiFe group 3c hydrogenases alongside HDR (heterodisulfide reductase) complex, suggesting a capacity for electron bifurcation and hydrogen cycling. The presence of a cytochrome heme oxidase complex and a putative nitrate reductase (NapA)-like gene hints at the potential for facultative anaerobic respiration.

Paralleling findings for Bathyarchaeia, the Thermoplasmatales genome encodes trimethylamine---corrinoid protein Co-methyltransferase (K14083), trimethylamine corrinoid protein (K14084), and dimethylamine corrinoid protein (K16179) indicate the capability for use of methylated compounds. Notably, the srvp_genome_38 genome encodes 31 copies of trimethylamine---corrinoid protein Co-methyltransferase, suggesting the high importance of methylated amines in the metabolic strategy of these soil-associated Thermoplasmatales archaea.

Only one representative (circularized) Methanomassiliicoccales genome contains genes for methyl-coenzyme M reductase (MCR), a key enzyme in methanogenesis. Other genes for methyl-dependent methane production were identified. The capacity to produce methane has been studied previously in this group, and it has been suggested that the capacity might have been laterally acquired (Zinke et al., 2020, Paul et al., 2012).

Nitrososphaeria

Nitrososphaeria comprises 4.16% of the wetland soil community. Genomes from the Nitrososphaerales order encode ammonia monooxygenase genes (K10944, K10945, K10946) and the nitrite reductase gene (K00368), indicating the capacity for ammonia oxidation and nitrite reduction. We also identified genes encoding lactate racemase (K22373) for interconverting L- and D-lactate, acylphosphatase (K01512) and acetyl-CoA synthetase (K01895) for acetate production, and propanol-preferring alcohol dehydrogenase (K13953), possibly for the production of alcohol.

The detection of genes encoding ornithine carbamoyltransferase (K00611), argininosuccinate synthase (K01940), argininosuccinate lyase (K01755) across several genomes from the Nitrososphaerales order, and the presence of carbamate kinase (K00926) and arginine deiminase (K01478) in at least one genome (srvp_genome_04), suggests the potential involvement of these archaea in the urea cycle or related ornithine-metabolic pathways.

One genome (srvp_genome_27) has genes involved in the AMP salvage pathway. AMP phosphorylase (K18931) breaks down AMP to produce ribose-1,5-bisphosphate; ribose 1,5-bisphosphate isomerase (K18237) converts ribose-1,5-bisphosphate to ribulose-1,5-bisphosphate and a Rubisco-like enzyme (K01601) incorporates CO2, forming glycerate 3P that can be fed into glycolysis or possibly, gluconeogenesis. We identified all the required genes for glycolysis except Pyruvate kinase isozymes R/L (K12406) and fructose-1,6-bisphosphatase I and 6-phosphofructokinase and 6-phosphofructokinase required for gluconeogenesis. Overall, we predict that Nitrososphaeria archaea in wetland soil contribute to nitrogen cycling involving ammonia, nitrate and urea, and carbohydrate transformations.

Hadearchaeia

We recovered 3 potentially complete genomes for Hadarchaeota. Previously, these archaea have been described from hydrothermal vents and hot springs where they are predicted to oxidize of alkanes. In the Hadearchaeales archaeal group, we identified genes for AMP phosphorylase (K18931), ribose 1,5-bisphosphate isomerase (K18237), and ribulose-bisphosphate carboxylase large chain (K01601), consistent with the AMP salvage pathway for formation of G3P.

In genomes from the Hadearchaeales group, we identified genes associated with one-carbon (C1) metabolism pathways, including methylenetetrahydrofolate dehydrogenase (NADP+) / methyltetrahydrofolate cyclohydrolase (K01491) and anaerobic carbon-monoxide dehydrogenase iron-sulfur subunit (K00196), along with 5,10-methylenetetrahydromethanopterin reductase (K00320). These genes may be involved in the reduction of methylene groups to methyl groups and the anaerobic oxidation of carbon monoxide. However, the absence of a complete set of genes for the Wood-Ljungdahl pathway (WLP) and the lack of genes for Acr-dependent alkane oxidation, which have been observed in other Hadarchaeota from deep-sea environments, suggests that these wetland soil Hadearchaeales may not utilize the full WLP for carbon fixation or the alkane oxidation pathway for energy conservation.

Hadearchaeales genomes encode genes for pyruvate ferredoxin oxidoreductase (porA K00169, porB K00170, porD K00171, porG K00172) and 2-oxoglutarate/2-oxoacid ferredoxin oxidoreductase (korA/oorA/oforA K00174, korB/oorB/oforB K00175, and an additional set of K00174 and K00175 for alpha and beta subunits, respectively, along with K00176 for the delta subunit), that likely enable oxidation of pyruvate and 2-oxoacids. The genomes encode fumarate hydratase subunits alpha (K01677) and beta (K01678) for conversion of malate to fumarate, but they lack a complete set of TCA cycle genes. This complex may be involved in amino acid biosynthesis.

The Hadearchaeales genomes exhibit complete pathways for the biosynthesis of coenzyme F420, a cofactor involved in various redox reactions, particularly in methanogens and some anaerobic bacteria. Their genomes also encode coenzyme F420 hydrogenase subunit beta (K00441), which enables the interconversion of reduced and oxidized versions of F420. The Hadearchaeales genomes encode NiFe hydrogenase 3b and two distinct groups of type 4 hydrogenases that can produce H_2.

Brockarchaeota

Brockarchaeota were found in three deep soil samples. There has only been one publication of the description of these archaea, and none were from soil (De Anda et al. 2021). One genome from 75 cm deep soil, srvp_genome_65, is potentially complete (complete pending resolution of one small polymorphism-rich area). In the Brockarchaeota genome srvp_genome_65, there was significant expression of genes encoding ABC transport system permease proteins (K02004). Genes for branched-chain amino acid transport system substrate-binding protein (K01999) were also highly expressed, highlighting the potential for importing amino acids. Among other genes expressed are the ribonucleoside-triphosphate reductase (K21636), Elongation factor 1-alpha (K03231) and DNA-directed RNA polymerase subunit B (K13798) involved in the synthesis of deoxyribonucleotides, protein synthesis and transcription, respectively. Interestingly, also expressed are the cell division protein FtsZ (K03531) and the archaeal cell division control protein

6 (K10725) indicating that they are actively replicating. The archaeal type IV pilus assembly protein PilA (K23255) is involved in cell adhesion and motility. The iron (III) transport system permease protein (K02011) is involved in iron uptake. Pyruvate ferredoxin oxidoreductase alpha subunit [EC:1.2.7.1] (K00169) is involved in the oxidation of pyruvate.

Consistent with the previously described genomes obtained from geothermally active environments, the Brockarchaeota-65 genome described in this study, encodes a relatively high number of carbohydrate-active enzymes (CAZY) compared to other soil archaea. This repertoire includes enzymes for the degradation of xylans, pectin, chitin, glucans, cellulose, and starch, indicating the ability to degrade carbohydrates with a wide range in complexity. Additionally, these archaea possess pathways for glycolysis/gluconeogenesis, serine to acetyl-serine conversion using AcCoA and H_2S, trimethylamine assimilation, and the reductive glycine pathway (rGlyP). Notably, acetate can be assimilated back to acetyl-CoA by an acetyl-CoA synthetase (ACS) reaction, which may be reversible given a [NiFe]-hydrogenase group 3c fused with an HDR complex.

Despite lacking the MCR complex for methane production, Brockarchaeota have genes for converting methanol, trimethylamine, and dimethylamine to Coenzyme M (or similar) and for converting CoM-S-S-Cob to Coenzyme B. They also appear capable of synthesizing the essential methanogenesis cofactor coenzyme F420 from a riboflavin-derived precursor, which is likely required for the 3c-Hdr complex and THMP cycling in reactions interconverting formate and serine. This metabolic profile is reminiscent of other archaeal groups such Bathyarchaeia, Freyarchaeota, and Atabeyarchaeota, which possess many genes associated with methanogenesis but are not classified as methanogens.

Brockarchaeota are likely involve in cycling nitrogen compounds in wetland soil, given genes for converting nitrate to NO (NirK) and nitrate to ammonia (NirBD), as well as potential pathways for converting formamide to ammonia. They can also metabolize dimethylsulfone and methylsulfonate to sulfite, producing an aldehyde and oxidizing FMN. These reactions typically require oxygen, which poses an interesting question regarding their occurrence in the oxygen-limited environment of deep wetland soils. However, the capability to oxidize FMNH could facilitate the regeneration of FMN in hydrogenase reactions, suggesting an adaptation to their anaerobic habitat.

Methanomethylicia

The complete genomes of Methanomethylicia, obtained from both short-read and long-read sequencing, reveal genes encoding methyl-coenzyme M reductase (MCR), but the complex is distinct from those of present in Methanomicrobiales and Methanosarcinales, thus we cannot determine if it is involved in methylotrophic methanogenesis or alkane metabolism. The srvp_genome_110 exhibits transcripts for MCR possibly signifying active methanogenesis/alkane metabolism. Additionally, the activity of genes encoding the F420-non-reducing hydrogenase complex (large subunit K14126, iron-sulfur subunit K14127, small subunit K14128) and the heterodisulfide reductase complex (HdrABC - subunit A2 K03388, subunit B2 K03389, subunit C2 K03390) indicates in situ reduction of heterodisulfide to ferredoxin, enabling energy conservation during methane or alkane metabolism.

Key to Methanomethylicia's metabolic capabilities is the presence of genes encoding the F420-non-reducing hydrogenase complex (K14126 large subunit, K14127 iron-sulfur subunit, K14128 small subunit) and the heterodisulfide reductase complex (HdrABC - K03388 subunit A2, K03389 subunit B2, K03390 subunit C2). These enzymes are involved in the exergonic, hydrogen-dependent reduction of heterodisulfide to ferredoxin, a critical step in energy conservation and electron transfer (e.g., during methanogenesis/alkane metabolism). Additionally, these genomes encode components of the nucleotide salvage pathway, as described above. Genes such as glycerol kinase (K00864), and glycerol-3-phosphate dehydrogenase (K00111) suggest Methanomethylicia's capability to utilize glycerol. This pathway supports the conversion of glycerol to glycerol-3-phosphate and further into triose phosphates, integrating into central carbon metabolism pathways. Overall, metabolic profiling of Methanomethylicia genomes suggests roles in the soil ecosystem related to either a methane production or alkane metabolism, nucleotide salvage, carbon fixation, and glycerol utilization.

Methanomicrobiales and methanotrophic archaea

The methanogenic archaea are the organisms that are most transcriptionally active in the soil samples. For Methanomicrobiales representative (srvp_genome-33), the most highly transcribed genes encode the methyl-coenzyme M reductase complex subunits. This finding confirms active methane production in the deep soil.

Also highly expressed are genes for transport of substrates necessary for methanogenesis, such as the formate (K21993). The presence of multiple subunits of the tetrahydromethanopterin S-methyltransferase complex (K00581 subunit E, K00579 subunit C, K00580 subunit D, K00577 subunit A, K00584 subunit H, K00582 subunit F), along with F420-reducing hydrogenase components (K00443 subunit gamma, K00440 subunit alpha, K00441 subunit beta, K00442 subunit delta), likely enable transfer of methyl groups and activity of reducing coenzyme F420, both of which may contribute to methanogenesis. Additionally, the presence of formate dehydrogenase subunits (K00123, K22516, K00125) suggests the utilization of formate as an electron donor in the methanogenic process. Transcription of the cell division protein FtsZ (K03531) and tubulin-like protein CetZ (K22222) indicate in situ growth and cell division. The identification of proteasome regulatory subunit (K03420) and transcription initiation factor (K03124) among the highly expressed genes points to the regulation of protein degradation and gene expression, essential for responding to environmental changes. Also expressed and important for growth are DNA-directed RNA polymerase subunits (K03045 subunit B, K03042 subunit A", K03041 subunit A'), elongation factor (K03231), and ribosomal proteins (K02867 L11). The heterodisulfide reductase (K03388 subunit A2), involved in recycling reduced coenzymes, is also expressed, as are V/A-type H+/Na+-transporting ATPase components (K02118 subunit B, K02117 subunit A, K02120 subunit D) that are crucial for maintaining proton or sodium gradients for ATP synthesis.

Overall, the transcriptomic data indicate that Methanomicrobiales are active in methane production and carbon cycling in wetland soils, utilizing formate, and hydrogen as substrates. The archaea are likely important sources of greenhouse gas emissions from the wetland soil.

Some Methanosarcinales (4.77% of the archaeal community) are methanotrophs that are known to be involved in the oxidation of methane (ref). In *Methanoperedens*_genome_01, we detected high expression levels of expression of genes for the methyl-coenzyme M reductase complex that is central to methane oxidation via the reverse methanogenesis pathway. Genes for nutrient and metal ion uptake (e.g., peptide/nickel transport system substrate-binding protein K02035) and for nucleotide metabolism (RNA-directed DNA polymerase, K00986) were also highly expressed.

Aenigmarchaeota

We recovered complete genomes from three major phyla of the DPANN archaea superphylum (Aenigmarchaeota, Woesearchaeota, and Micrarchaeota) from the SRVP soil metagenomes and metatranscriptomes. Transcripts were only recovered from a representative genome of Aenigmarchaeota. For Aenigmarchaeota-80, expressed genes encoded putative membrane proteins (K08982), ribose-phosphate pyrophosphokinase (K00948) involved in the nucleotide salvage pathway, and enzymes such as the 5-methylcytosine-specific restriction enzyme (K07451). This suggests the importance of DNA repair, modification, and transcription regulation. The expression of genes involved in protein synthesis (e.g., large subunit ribosomal proteins K02912) and DNA-directed RNA polymerase subunit F (K03051) indicate in situ metabolic activity. Among the highly transcribed hypothetical genes are S-layer proteins, Given that DPANN archaea are widely predicted to be episymbionts that probably only grow and replicate when attached to host cells (e.g., *Nanobdella aerobiophila* that associate with methanogen (Kato et al. 2022), these findings suggest that this Aenigmarchaeota is an episymbionts of an archaeal host in the wetland soil, possibly coexisting methanogens.

Aenigmarchaeota representative genomes and associated MAGs have expressed nucleoside-5'-monophosphate phosphorylase (NMP phosphorylase), previously known as AMP phosphorylase, which is involved in the phosphorolysis of AMP, CMP, and UMP (Sato et al. 2022). This enzyme catalyzes the breakdown of these nucleoside monophosphates in the nucleotide salvage pathway. The genomes also encode NiFe-hydrogenases, with one group lacking the 3b hydrogenase and instead having an unclassified group 4 hydrogenase. Aenigmarchaeota exhibits a partial glycolysis pathway, and only three representative MAGs have fructose 1,6-bisphosphate aldolase/phosphatase (K01622), a bifunctional enzyme with both aldolase and phosphatase activities (Say and Fuchs 2010). This enzyme is considered an ancestral gluconeogenic enzyme present in various archaeal groups and deeply branching bacterial lineages, indicating its importance in metabolic pathways across different organisms.

Micrarchaeota

The expression of genes related to isoprenoid synthesis, DNA replication, protein glycosylation, and the presence of an S-layer in the Micrarchaeota MAG is consistent with previous findings on DPANN archaea, despite their putative episymbiotic lifestyle. The presence of the phosphomevalonate kinase gene (K00938) suggests that Micrarchaeota may be capable of synthesizing isoprenoids or sterols via the mevalonate pathway. This is in line with previous studies that have identified genes involved in lipid biosynthesis in DPANN archaea (Li et al. 2021; Chen et al. 2018). The expression of the DNA primase gene (K02316) indicates active DNA replication processes, which is essential for the survival and propagation of these organisms, regardless of their symbiotic relationships. The expression of the oligosaccharyl transferase STT3 subunit gene (PF02516.16) suggests that Micrarchaeota may perform N-linked glycosylation. This post-translational modification has been observed in other archaea and is important for protein stability and function. Lastly, the presence of a gene encoding a protein with an S-layer-like family N-terminal region (PF05123.14) confirmed that Micrarchaeota have an S-layer, common cell surface structure in archaea.

5.4 Conclusion

The application of PacBio Revio HiFi sequencing and the new hifiasm-meta and metaMDBG assembly algorithms enabled the reconstruction of over 1,200 metagenome-assembled genomes (MAGs) from wetland soil samples. Of these > 500 are circularized and/or complete. This is an important advance, technically, because soils are microbially extremely complex so, to date, very few circularized/complete genomes have been reported from soil of any type. The availability of complete genomes allowed for exact measures of genome size, prediction of the replication mode, coding structure, gene content, and precise linkage between RNA transcripts and their source organisms. The 110 circularized/ complete archaeal genomes span evolutionarily divergent groups such as Bathyarchaeia, Asgardarchaeota and DPANN archaea and, in combination with transcriptomic data, provided evidence of *in situ* activity in deep (>60 cm), anaerobic wetland soil. We uncovered an intertwined set of metabolisms that includes methane production supported by microbial generation of small organic substrates and H_2 (generated via widely distributed hydrogenases). Countering methane emissions from this wetland soil are coexisting archaea that are predicted to generate energy via oxidation of a subset of the H_2 pool and coexisting methanotrophs that also have the capacity to fix N_2 into ammonia. Ammonia oxidation is mediated by other Thaumarchaea, which contribute to other aspects of the soil nitrogen cycle. Overall, carbon substrates, probably mostly microbial detritus in the form of fatty acids, polysaccharides and amino acids and their breakdown products, as well as CO_2 fixation represent the important substrates supporting in situ activity. The presence of genes for methylated compound metabolism encoded in the genomes of several archaea suggests the importance of compounds such as methylamines in soil as carbon/nitrogen sources whereas breakdown of sulfonated carbon compounds likely provides a source of reduced sulfur.

5.5 Figures

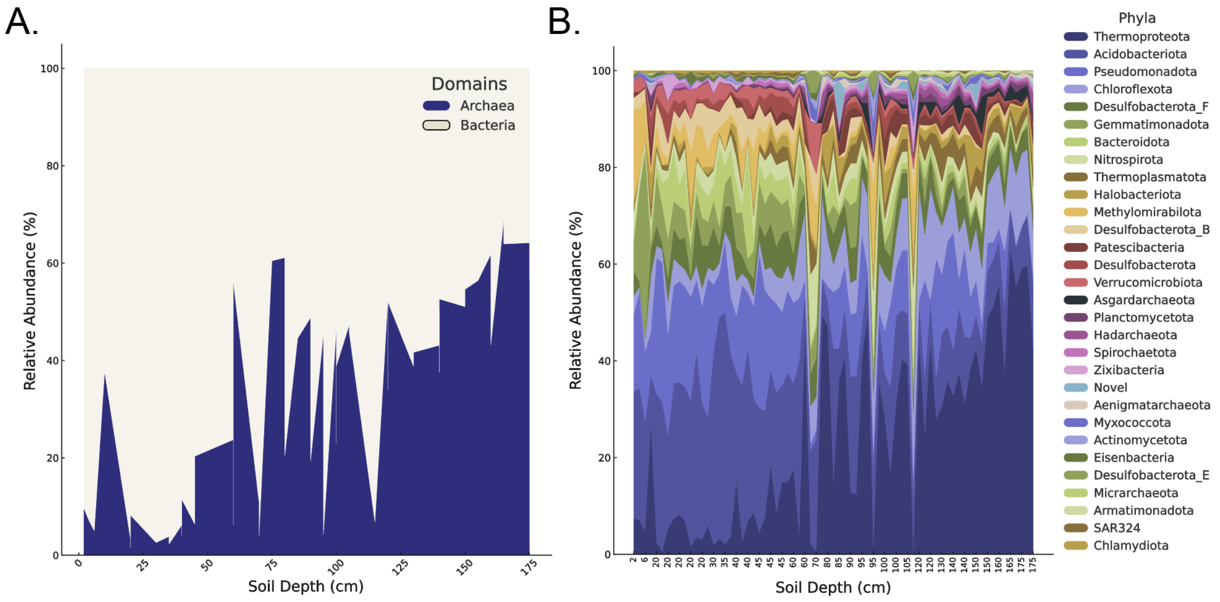


Figure 5.1 Relative Abundance of Microbial Domains by Soil Depth based on MAGs based on an analysis of all available genomes. A) Stacked bar chart illustrating relative abundances of Bacteria and Archaea as a function of the soil depth in centimeters. B) Distribution of archaeal and bacteria phyla based on GTDB as a function of the soil depth profile.

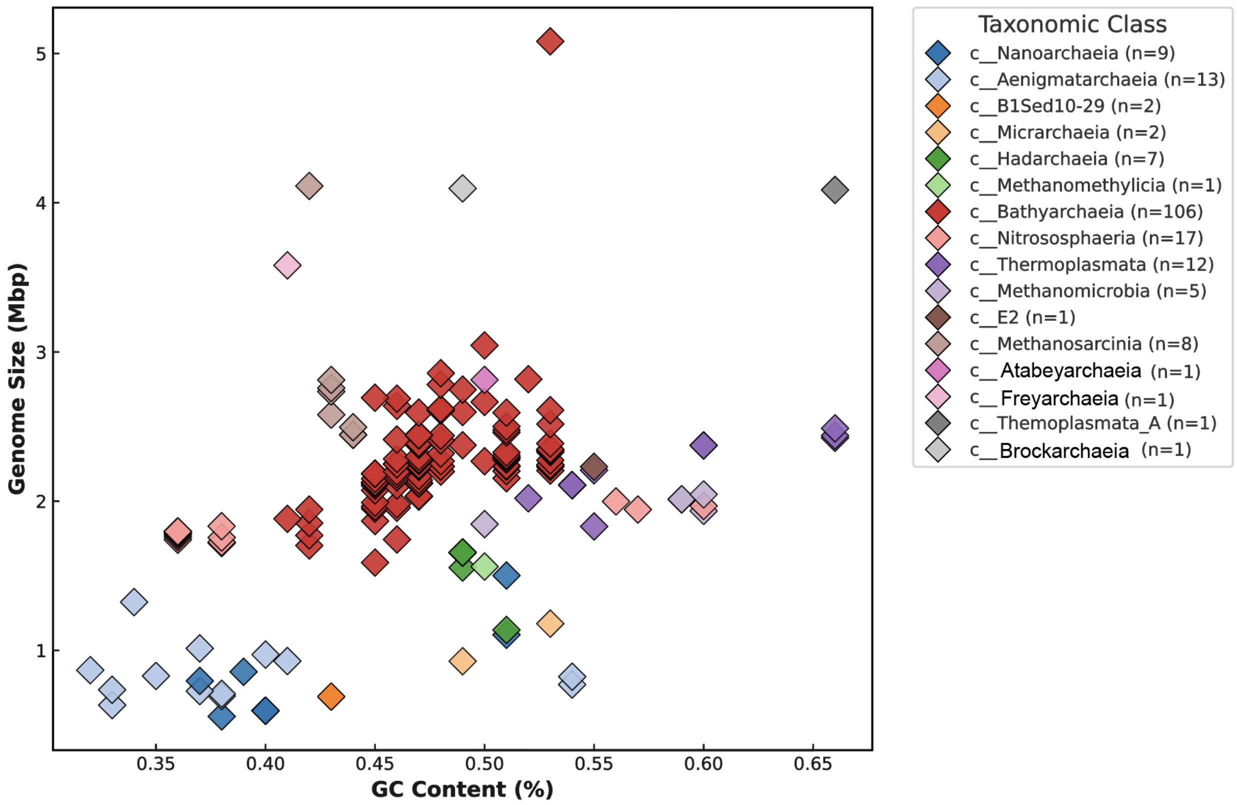


Figure 5.2 Genome Size versus GC Content for High-Quality Circular Genomes Assembled and Dereplicated at 95% from PacBio Sequencing. This scatter plot visualizes the relationship between genome size (megabase pairs, Mbp) and GC content (%) for high-quality circular genomes assembled from PacBio sequencing data, categorized by their taxonomic class as defined by the Genome Taxonomy Database (GTDB). Each diamond marker represents a single genome, with its position indicating the genome's GC content and size. The plot highlights the variation in genome size and GC content across various microbial taxa, underscoring the effectiveness of long-read sequencing in capturing a wide range of genomic architectures. This analysis reveals patterns of genomic variation that may reflect ecological adaptations and evolutionary trajectories within microbial communities.

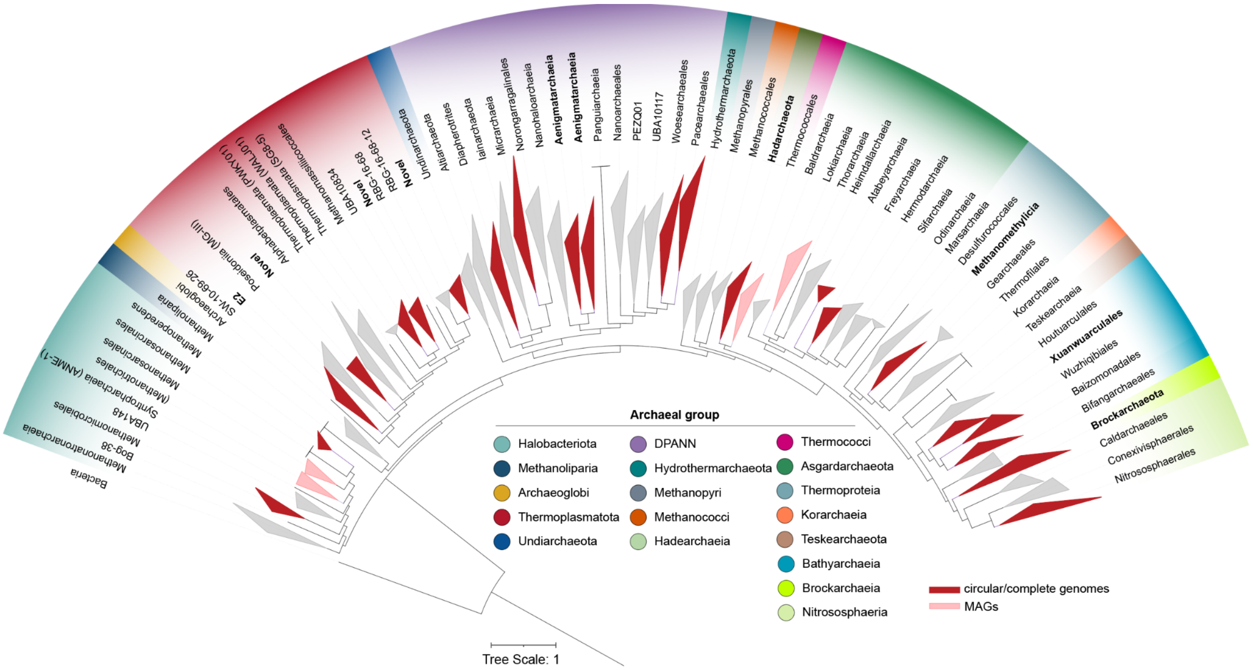


Figure 5.3 Phylogenetic tree constructed from concatenated ribosomal protein 15 (RP15) sequences, showing the archaeal taxonomic diversity in SRVP wetland soil. We used 969 archaeal reference genomes and 118 SRVP wetland soil high quality genomes.

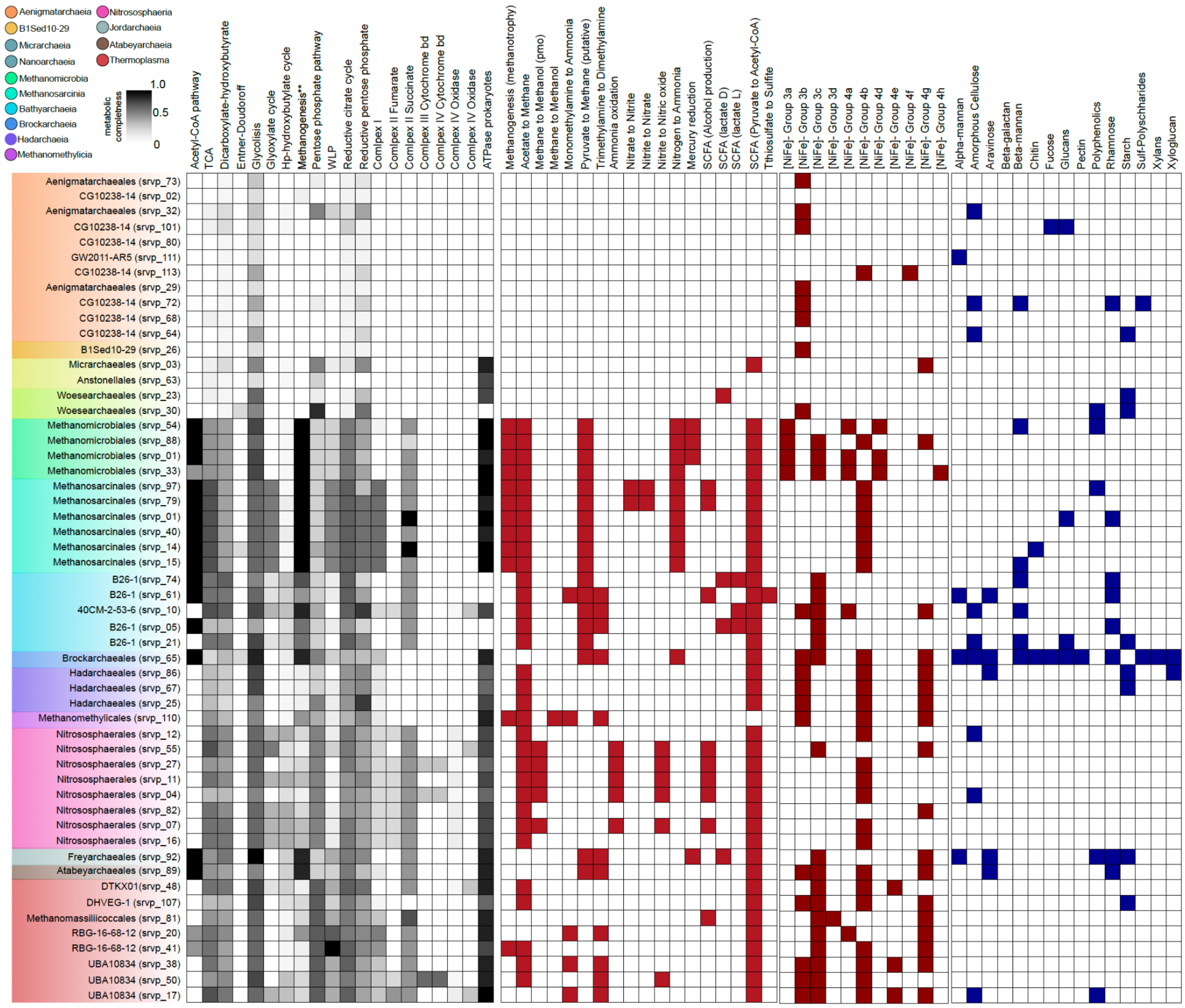


Figure 5.4 Heatmap showing presence of metabolic genes in circular complete Archaeal genomes involved in energy, carbohydrates metabolism, nitrogen, hydrogen and carbon assimilation. On the left panel, genomes organized by taxonomic affiliation (dereplicated >95% completeness, <5% contamination). In parentheses are the names of the short names of the genomes, Table X has full information about the genomes, accession numbers and GTDB classification.

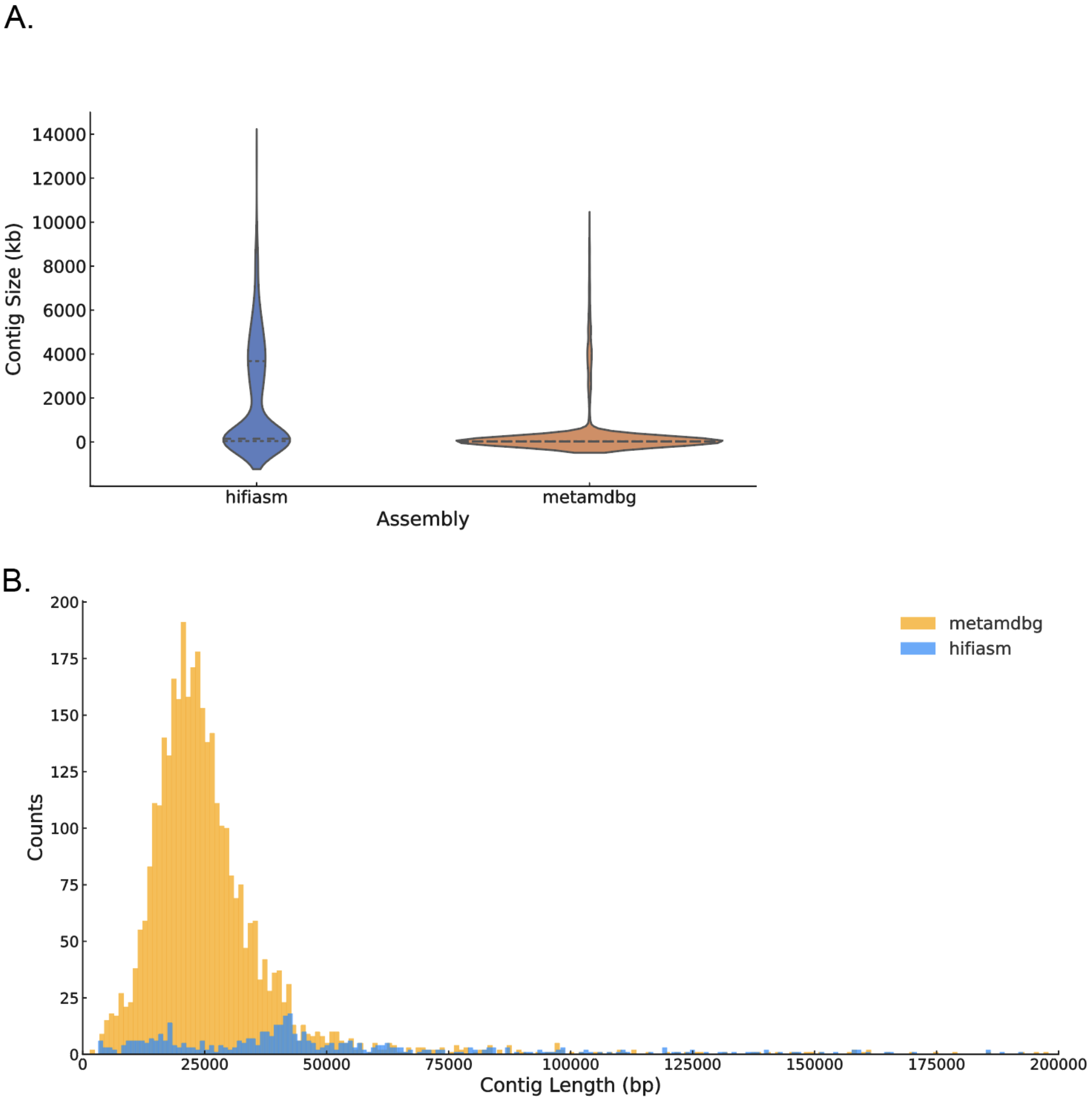


Figure S1 A) Size distribution of contigs for hifiasm and metamdbg assemblies. B) Distribution of circular contigs <200 kilobases. Histogram plot displaying the distribution of circular contigs smaller than 200 kb for both hifiasm and metamdbg assemblies. The contig lengths are plotted along the x-axis ranging from 0 to 200,000 base pairs (bp), and the y-axis shows the count of contigs within each length interva

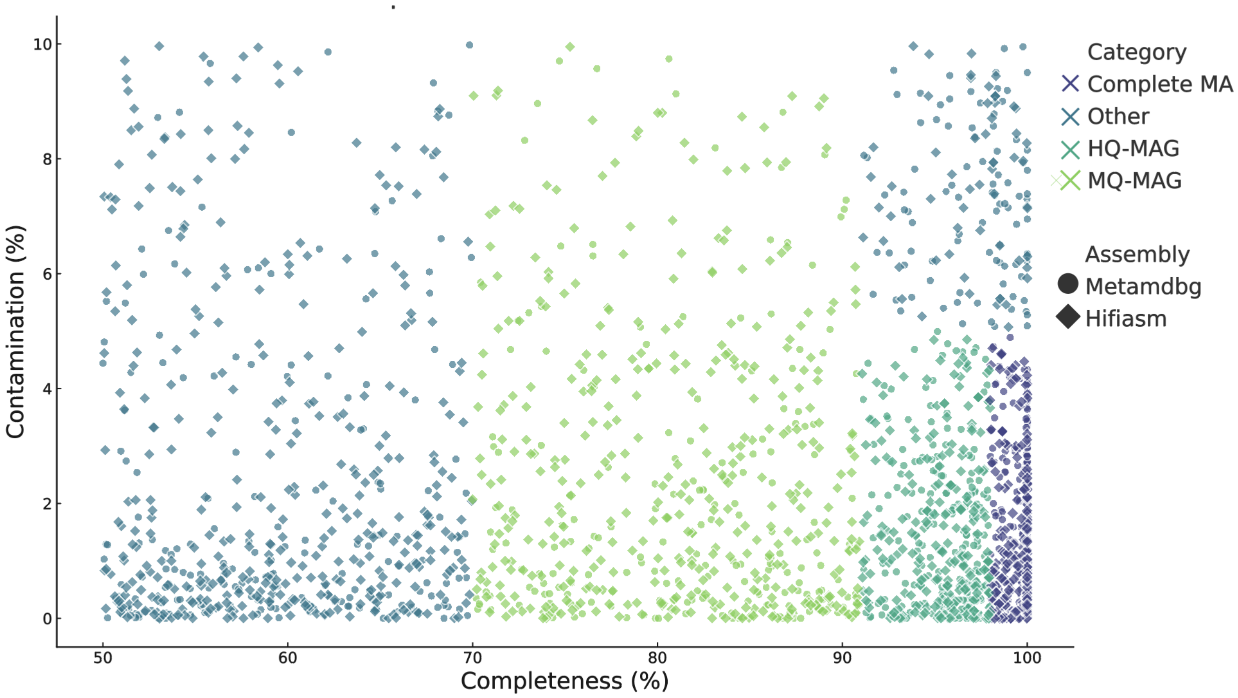


Figure S2 Comparative quality assessment of metaMDBG and hifiasm-meta Assembly Outputs. The scatter plot delineates the nuanced interplay between completeness and contamination metrics across MAGs procured via metaMDBG and hifiasm-meta assembly protocols. Distinct markers symbolize the two methodologies: circles for metaMDBG and diamonds for hifiasm-meta, facilitating an intuitive comparison. The MAGs are stratified into predefined quality tiers based on their genomic completeness and contamination based on single copy genes, coverage and checkM2. Complete MAGs are epitomized by 98-100% completeness alongside a contamination threshold below 5%. High-Quality MAGs (HQ-MAGs) are characterized by 91-97% completeness with contamination maintained under 5%. Medium-Quality MAGs (MQ-MAGs) span the 70-90% completeness range, with a permissible contamination ceiling of 10%. Low-Quality MAGs reflect a compromised genomic representation, falling below 50% completeness and surpassing 10% contamination.

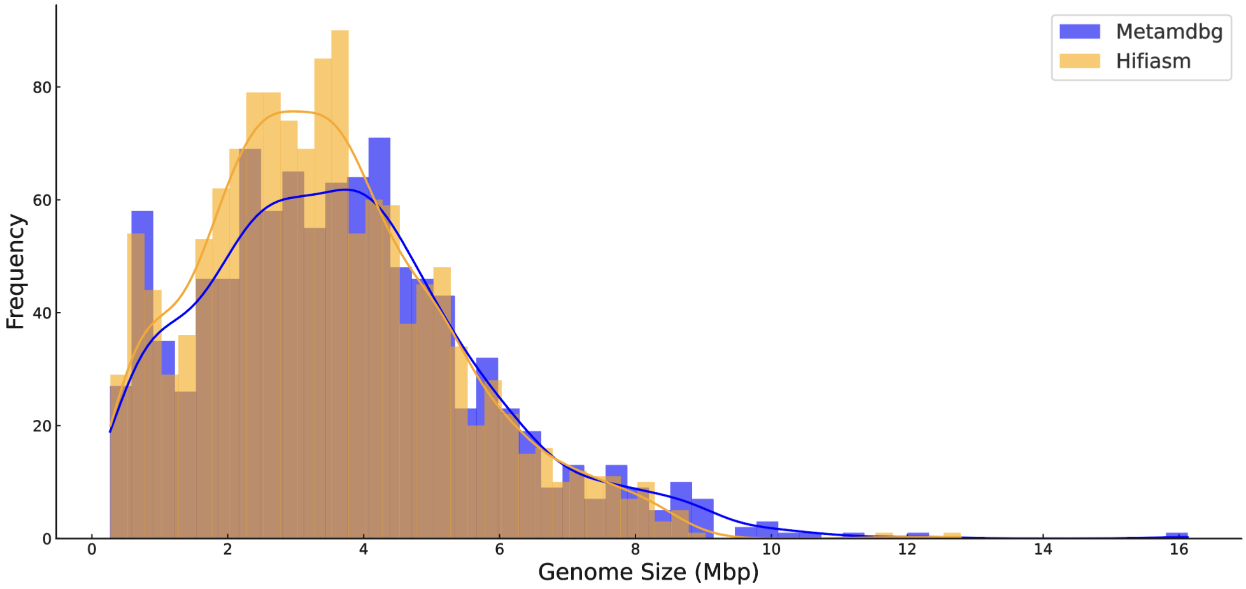


Figure S3 Comparative distribution of genome sizes for metaMDBG and hifiasm-meta Assemblies. Distribution of genome sizes among the MAGs recovered from the metaMDBG and hifiasm-meta assemblies, represented in megabase pairs (Mbp). The blue histogram represents the metaMDBG assembly, while the orange histogram represents the hifiasm-meta assembly. Both histograms include a kernel density estimate (KDE) to illustrate the overall shape of the genome size distribution. This analysis highlights the diversity in genome sizes captured by each assembly method, with both assemblies exhibiting a broad range of sizes.

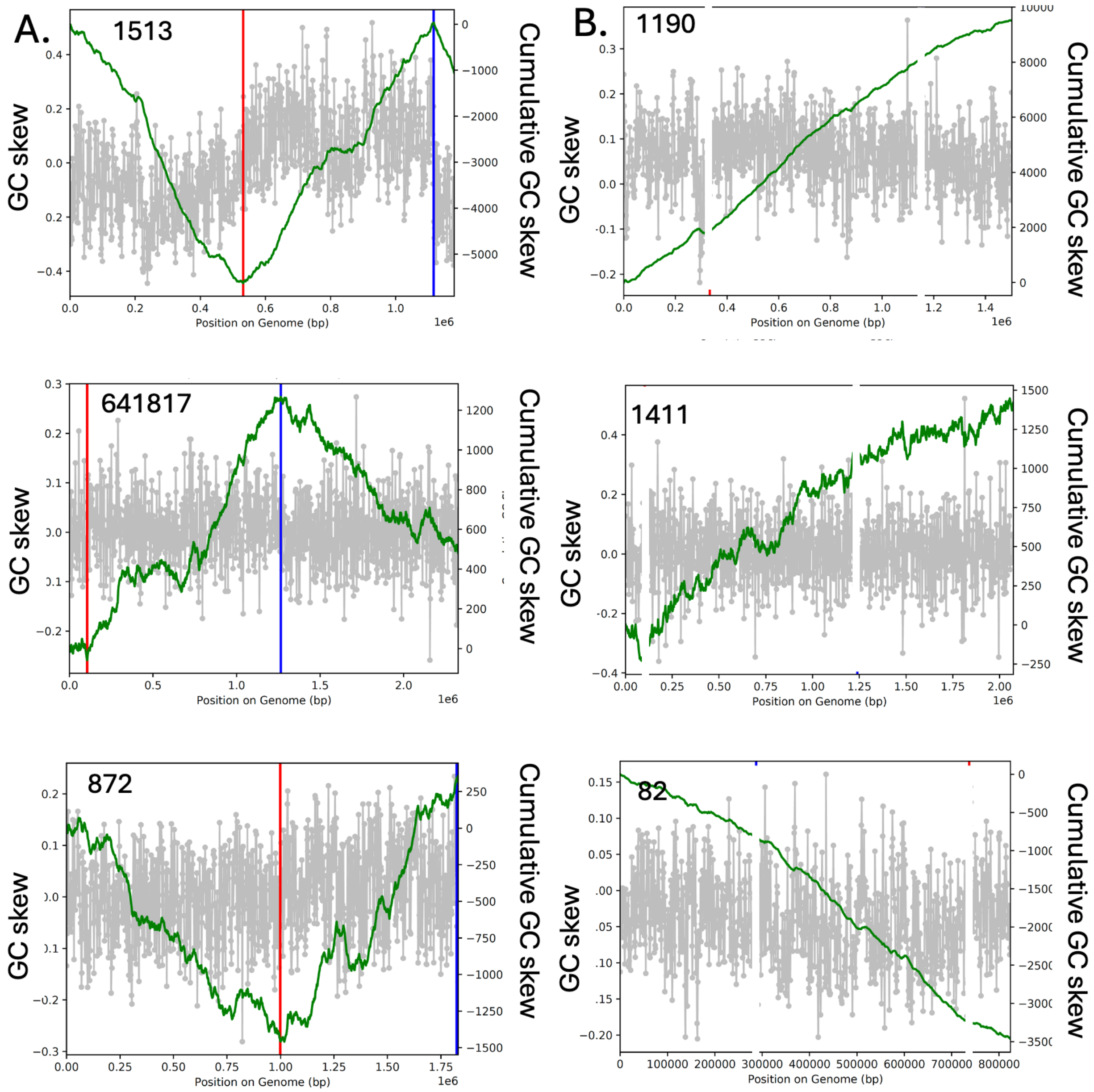
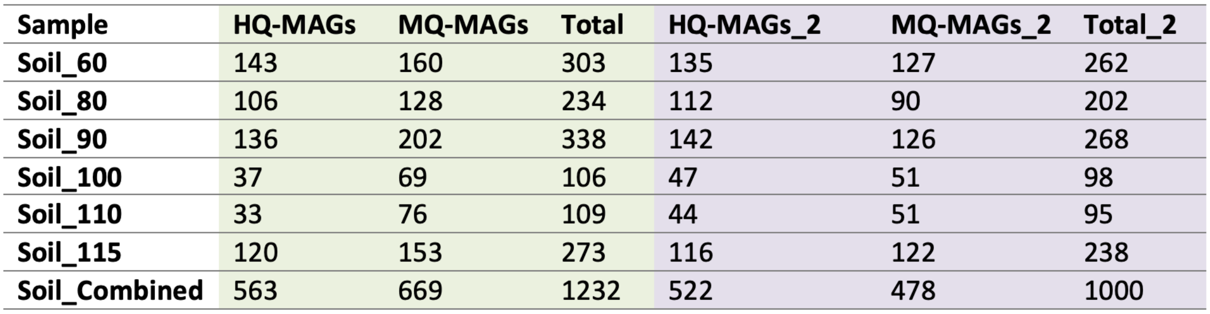


Figure S4 GC skew of 6 potentially complete genomes from PacBio dataset. The GC skew is shown as a gray plot and the cumulative GC skew is overlain (green line).

5.6 Tables



Sample	HQ-MAGs	MQ-MAGs	Total	HQ-MAGs_2	MQ-MAGs_2	Total_2
Soil_60	143	160	303	135	127	262
Soil_80	106	128	234	112	90	202
Soil_90	136	202	338	142	126	268
Soil_100	37	69	106	47	51	98
Soil_110	33	76	109	44	51	95
Soil_115	120	153	273	116	122	238
Soil_Combined	563	669	1232	522	478	1000

Supplementary Table 1 Comparison of the assembly performance of Hifiasm-meta (olive color) and metaMDBG (lilac).

5.7 References

Al-Shayeb, B., Schoelmerich, M. C., West-Roberts, J., Valentin-Alvarado, L. E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M. J., Williams, K. H., Doudna, J. A., & Banfield, J. F. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature*, *610*(7933), 731–736.

Angle, J. C., Morin, T. H., Solden, L. M., Narrowe, A. B., Smith, G. J., Borton, M. A., Rey-Sanchez, C., Daly, R. A., Mirfenderesgi, G., Hoyt, D. W., Riley, W. J., Miller, C. S., Bohrer, G., & Wrighton, K. C. (2017). Methanogenesis in oxygenated soils is a substantial fraction of wetland methane emissions. *Nature Communications*, *8*(1), 1567.

Benoit, G., Raguideau, S., James, R., Phillippy, A. M., Chikhi, R., & Quince, C. (2024). High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01983-6

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S. T., Shin, S. B., Zorea, A., Andreu, V. P., Panke-Buisse, K., Medema, M. H., Mizrahi, I., Pevzner, P. A., & Smith, T. P. L. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*, *40*(5), 711–719.

Blohs, M., Moissl-Eichinger, C., Mahnert, A., Spang, A., Dombrowski, N., Krupovic, M., & Klingl, A. (2019). Archaea – An Introduction. In T. M. Schmidt (Ed.), *Encyclopedia of Microbiology (Fourth Edition)* (pp. 243–252). Academic Press.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211.

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/servlets/purl/1241166

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., & Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews. Microbiology*, *16*(10), 629–645.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., & Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315–333.

Chen, L.-X., Méndez-García, C., Dombrowski, N., Servín-Garcidueñas, L. E., Eloe-Fadrosh, E. A., Fang, B.-Z., Luo, Z.-H., Tan, S., Zhi, X.-Y., Hua, Z.-S., Martinez-Romero, E., Woyke, T., Huang, L.-N., Sánchez, J., Peláez, A. I., Ferrer, M., Baker, B. J., & Shu, W.-S. (2018). Metabolic versatility of small archaea Micrarchaeota and Parvarchaeota. *The ISME Journal*, *12*(3), 756–775.

De Anda, V., Chen, L.-X., Dombrowski, N., Hua, Z.-S., Jiang, H.-C., Banfield, J. F., Li, W.-J., & Baker, B. J. (2021). Brockarchaeota, a novel archaeal phylum with unique and versatile carbon cycling pathways. *Nature Communications*, *12*(1), 2404.

Dombrowski, N., Williams, T. A., Sun, J., Woodcroft, B. J., Lee, J.-H., Minh, B. Q., Rinke, C., & Spang, A. (2020). Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nature Communications*, *11*(1), 3939.

Eisenhofer, R., Odriozola, I., & Alberdi, A. (2023). Impact of microbial genome completeness on metagenomic functional inference. *ISME Communications*, *3*(1), 12.

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., Seitz, K. W., Dombrowski, N., Lewis, W. H., Homa, F., Saw, J. H., Lombard, J., Nunoura, T., Li, W.-J., Hua, Z.-S., Chen, L.-X., Banfield, J. F., John, E. S., Reysenbach, A.-L., … Ettema, T. J. G. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, *618*(7967), 992–999.

Evans, P. N., Boyd, J. A., Leu, A. O., Woodcroft, B. J., Parks, D. H., Hugenholtz, P., & Tyson, G. W. (2019). An evolving view of methane metabolism in the Archaea. *Nature Reviews. Microbiology*, *17*(4), 219–232.

Feng, X., Cheng, H., Portik, D., & Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, *19*(6), 671–674.

Guerra, C. A., Heintz-Buschart, A., Sikorski, J., Chatzinotas, A., Guerrero-Ramírez, N., Cesarz, S., Beaumelle, L., Rillig, M. C., Maestre, F. T., Delgado-Baquerizo, M., Buscot, F., Overmann, J., Patoine, G., Phillips, H. R. P., Winter, M., Wubet, T., Küsel, K., Bardgett, R. D., Cameron, E. K., … Eisenhauer, N. (2020). Blind spots in global soil biodiversity and ecosystem function research. *Nature Communications*, *11*(1), 3870.

He, C., Keren, R., Whittaker, M. L., Farag, I. F., Doudna, J. A., Cate, J. H. D., & Banfield, J. F. (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nature Microbiology*, *6*(3), 354–365.

Kato, S., Ogasawara, A., Itoh, T., Sakai, H. D., Shimizu, M., Yuki, M., Kaneko, M., Takashina, T., & Ohkuma, M. (2022). Nanobdella aerobiophila gen. nov., sp. nov., a thermoacidophilic, obligate ectosymbiotic archaeon, and proposal of Nanobdellaceae fam. nov., Nanobdellales ord. nov. and Nanobdellia class. nov. *International Journal of Systematic and Evolutionary Microbiology*, *72*(8). https://doi.org/10.1099/ijsem.0.005489

Kayranli, B., Scholz, M., Mustafa, A., & Hedmark, Å. (2010). Carbon Storage and Fluxes within Freshwater Wetlands: a Critical Review. *Wetlands*, *30*(1), 111–124.

Li, Y.-X., Rao, Y.-Z., Qi, Y.-L., Qu, Y.-N., Chen, Y.-T., Jiao, J.-Y., Shu, W.-S., Jiang, H., Hedlund, B. P., Hua, Z.-S., & Li, W.-J. (2021). Deciphering Symbiotic Interactions of “Candidatus Aenigmarchaeota” with Inferred Horizontal Gene Transfers and Co-occurrence Networks. *mSystems*, *6*(4), e0060621.

Oliverio, A. M., Narrowe, A. B., Villa, J. A., Rinke, C., Hoyt, D. W., Liu, P., McGivern, B. B., Bechtold, E. K., Ellenbogen, J. B., Daly, R. A., Smith, G. J., Angle, J. C., Flynn, R. M., Freiburger, A. P., Louie, K. B., Stemple, B., Northen, T. R., Henry, C., Miller, C. S., … Wrighton, K. C. (2024). Rendering the metabolic wiring powering wetland soil methane production. In *bioRxiv* (p. 2024.02.06.579222). https://doi.org/10.1101/2024.02.06.579222

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), 833–844.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437.

Sato, T., Utashima, S. H., Yoshii, Y., Hirata, K., Kanda, S., Onoda, Y., Jin, J.-Q., Xiao, S., Minami, R., Fukushima, H., Noguchi, A., Manabe, Y., Fukase, K., & Atomi, H. (2022).

A non-carboxylating pentose bisphosphate pathway in halophilic archaea. *Communications Biology*, *5*(1), 1290.

Say, R. F., & Fuchs, G. (2010). Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature*, *464*(7291), 1077–1081.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*, *48*(16), 8883–8900.

Søndergaard, D., Pedersen, C. N. S., & Greening, C. (2016). HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports*, *6*, 34212.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37–43.

Valentin-Alvarado, L. E., Appler, K. E., De Anda, V., Schoelmerich, M. C., West-Roberts, J., Kivenson, V., Crits-Christoph, A., Ly, L., Sachdeva, R., Savage, D. F., Baker, B. J., & Banfield, J. F. (2023). Asgard archaea modulate potential methanogenesis substrates in wetland soil. In *bioRxiv* (p. 2023.11.21.568159). https://doi.org/10.1101/2023.11.21.568159

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., & Anantharaman, K. (2022). METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, *10*(1), 33.

Concluding remarks and future work

This dissertation has made significant contributions to our understanding of the diversity, metabolic capabilities, and evolutionary dynamics of enigmatic microbial lineages, particularly those found in subsurface environments. By leveraging genome-resolved metagenomics, bioinformatics, and structural biology, we have revealed novel insights into the roles of these microorganisms in biogeochemical cycles, their symbiotic relationships, and their potential impact on climate change.

The exploration of genomes from metagenomes has revolutionized our ability to discover and study novel microbial lineages that were previously inaccessible through cultivation-based approaches. A fundamental aspect of my doctoral research was the reconstruction of complete genomes from metagenomes, which provides a comprehensive view of an organism's genetic repertoire and enables a deeper understanding of its evolutionary history, metabolic capabilities, and ecological roles.

Our findings have revealed the presence of active and ancient lineages of [FeFe]-hydrogenases in anaerobic archaea, including a novel ultra-minimal hydrogenase in DPANN archaea and remarkable hybrid complexes formed through the fusion of [FeFe]- and [NiFe]-hydrogenases. These discoveries not only shed light on new metabolic adaptations of archaea but also provide streamlined H_2 catalysts for potential biotechnological applications.

Furthermore, we have identified the roles of soil-associated Asgard archaea in modulating potential methanogenesis substrates in wetland soils. Our metatranscriptomic analyses highlight the high expression of genes involved in hydrogen metabolism, carbon fixation, and the breakdown of various organic compounds. These findings suggest that Asgard archaea play a significant role in shaping the reservoirs of substrates for methane production in terrestrial ecosystems.

The exploration of mobile genetic elements (MGEs) in Asgard archaea has revealed the importance of defense systems and methylation mechanisms in modulating their interactions with MGEs. The integration, excision, and copy number variation of MGEs enable host genetic versatility, contributing to the ongoing evolution of these archaea.

While the recovery of near-complete metagenome-assembled genomes (MAGs) has been a significant advancement, the curation of complete genomes offers additional insights into the gain or loss of functions and the evolutionary trajectories of these organisms. With the advent of long-read sequencing technologies, such as PacBio and Nanopore, we can now recover complete genomes more efficiently, although these genomes still require manual validation and characterization of their genome structure and associated extrachromosomal elements, if present.

Lastly, the application of long-read sequencing technologies, particularly PacBio Revio HiFi reads, has greatly advanced our ability to recover complete microbial genomes from complex environments. The availability of complete genomes has enabled precise measures of genome size, architecture, gene content, and the linkage between RNA transcripts and their originating organisms. This has provided valuable insights into the structure, function, and gene expression patterns of microbial communities in wetland soils. This approach has significantly reduced the curation time and provided us with a wealth of finalized genomes, enabling us to comprehensively investigate the evolutionary, metabolic, and ecological histories of these organisms that have been challenging to study in the laboratory.

One of the significant challenges in modern genomics is the *in silico* prediction of hypothetical proteins, as most proteins in many organisms remain unknown. Traditional approaches, such as guilt-by-association, phylogenetic analysis, and experimental studies, can be time-consuming and laborious when studying hundreds of proteins simultaneously. To address this challenge, we employed the cutting-edge protein structure prediction tool AlphaFold, which has advanced the screening process for hundreds of proteins and provided a robust way to complement phylogenetic inferences.

By combining the power of complete genome reconstruction from metagenomes, phylogenetic analyses, and state-of-the-art structural prediction tools like AlphaFold, we have gained unprecedented insights into the diversity, evolution, and functional capabilities of previously understudied microbial lineages. This integrative approach has enabled us to unravel the complexities of these enigmatic organisms and their roles in various biogeochemical processes, paving the way for future discoveries and applications in biotechnology and environmental remediation.

This dissertation is merely a stepping stone, paving the way for future discoveries and applications in biotechnology, environmental microbiology, and our quest to unravel the intricate tapestry of life on Earth.