

Singled Out: Genomic analysis of uncultured microbes in marine sediments

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Jordan Toby Bird
December 2017**

Copyright © 2017 by Jordan Bird
All rights reserved.

Dedication

To my loving parents, Trent and Lori, and amazing and compassionate brother, Justin.

Acknowledgements

Firstly, this work would not have been possible without the support and guidance of my advisor, Karen G. Lloyd. She took a chance on a strong willed, challenging student to be her first graduate student. She went above and beyond her charge as an advisor, and her dogged pursuit of knowledge and endless generosity of spirit has been a welcomed constant in my life of the last five years. Moreover, she will forever on be a shining example of personhood in my life as her unwavering ability to see and address the humanity in every person she meets remains a model I will endeavor to emulate. Secondly, I would like to thank my friends and colleagues that have kept me grounded. A special thanks to those who have sat and listened to me during times of despair and to those who have celebrated with me in times of success. My fellow labmates—Joy, Richard, Katie, and Kate—have inspired me and endured me at my most outwardly cynical. Thanks to my SCALE-IT cohort and our leader Dr. Harry Richards. His pursuit to see us all succeed did result in some of the key learning experiences that I will carry with me through the next stage of my career. When I first visited the University of Tennessee, I noted the eagerness of the faculty to engage students, collaborate, and provide resources and ideas as a key part of my decision to enroll in the program. This observation has remained true throughout my years in Knoxville. To my Geobio peeps, you are always a source of inspiration for me, and the memories in science and community we shared are some of my fondest. Thanks to Dr. Brett Baker, Dr. Alexander Probst, and Dr. Brandi Reese for being wonderful and accommodating collaborators outside of UTK. To my family, whose support throughout my time here, but especially during the writing of this dissertation was crucial and extraordinary. I will forever be in debt to you.

Abstract

The vast majority of abundant taxa in marine sediment environments have not yielded to culture, leaving questions about their relationship to other taxa and their functional potential unanswered. However, in the absence of active cultures, careful application of various omics methods can be used to help us make useful inferences about their evolutionary history and how they have continued to survive in environments of extreme energy deprivation. For this dissertation, I have applied comparative genomics methods to members of two uncultured groups, the recently proposed Altiarchaeales order and a cosmopolitan taxon associated with the Actinobacteria phylum. Additionally, I combined transcript recruitment and metabolomic profiles to investigate metabolisms inferred from the single-cell amplified genomes extracted from members of a taxa that thrive in Baltic Sea sediment microbial communities. In Chapter II, I establish a phylogenetic relationship across distantly related members of the order Altiarchaeales and discuss environment-specific adaptations. In Chapter III, transcript recruitment and metabolite profiles support a community-wide focus on microbial persistence with active members of the uncultured Atribacteria phylum playing an important ecological role. In Chapter IV, my analysis leads to the proposal of the new class within the Actinobacteria. Osirisbacteria is a class of Actinobacteria that is specialized for life in anoxic environments. Overall, this work offers new insights into deeply-branching microbial taxa, improved understanding of recently considered branches of the evolutionary tree, and new perspective on metabolisms important for survival in low-energy marine sediment environments.

Table of Contents

Chapter I: Introduction.....	1
List of References	12
Chapter II: Culture independent genomic comparisons reveal environmental adaptations for Altiarchaeales	20
Abstract	22
Introduction	23
Materials and Methods	26
Results	32
Discussion	44
Conclusion.....	50
Funding	50
Acknowledgements.....	51
List of References	52
Chapter III: Atribacteria support community-wide microbial persistence through 44,000 years of Baltic Sea sedimentation	59
Abstract	61
Main Text.....	62
Materials and Methods	77
Acknowledgements.....	81
List of References	82
Chapter IV: Comparative genomic analysis reveals earliest branching Actinobacteria class, Candidatus Osiribacteria	86
Abstract	88
Introduction	88
Results and Discussion	89
Materials and Methods	104
Acknowledgements.....	106
List of References	107
Chapter V: Conclusion	115

List of References	119
Vita	121

List of Tables

Table 2.1	Statistics of genomes highlighted in this study	36
Table 2.2	Shared replication machinery across Altiarchaeales genomes	48
Table 3.1	Genome sources and accessions	66
Table S3.1	Enzyme substrates and nominal corresponding enzymes	85
Table 4.1	Level of overlap between orthologs in Actinobacteria OPB41	98

List of Figures

Figure 1.1	Lloyd Lab at the White Oak River Estuary in Fall of 2013	3
Figure 1.2	 Picture of R/V Greatship Manasha using during IODP Expedition 347.....	4
Figure 1.3	Schematic of single-cell genomics and metagenomic reconstruction techniques	5
Figure 1.4	Map of sample sites from IODP Expedition.....	9
Figure 2.1	Map of global distribution of Altiarchaeales 16S rRNA gene sequences	25
Figure 2.2	Graph of aqueous porewater concentrations of sulfate and methane in cores taken from White Oak River Estuary.....	28
Figure 2.3	Map of sample sites from IODP Expedition.....	37
Figure 2.4A	Maximum likelihood tree based on full length 16S rRNA gene sequences	38
Figure 2.4B	Phylobayes tree based on 10 conserved proteins from 94 archeal genomes...	39
Figure 2.5	Alignment of amino acids of pheT sequences in six proposed Altiarchaeales with representative pheT sequences.....	40
Figure 2.6	Annotated alignment showing mismatches in common full length archeal primers	42
Figure 2.7	Comparative analysis of protein clusters within all known Altiarchaeales genomes	43
Figure 3.1	Single amplified genomes from diverse and abundant bacterial lineages	64
Figure 3.2	Operational taxonomic units (OTUs) composition for three 16S rRNA amplicon libraries and SAGs recovered from Baltic sea sediment horizons	65
Figure 3.3	Downcore profiles of each small organic molecule detected in the Baltic Sea sediment	68
Figure 3.4A	Transcriptional and translational regulatory elements found in transcriptomes of bacterial SAGS	70
Figure 3.4B	Proportions of transcriptional and translational regulatory elements also found in the metatranscriptomes	71
Figure 3.5	Correlations between enzyme activities and mapped transcript abundances for relevant genes in M0059	73
Figure 3.6	Diagram of key metabolic pathways of Atribacteria.....	75

Figure 4.1	A maximum-likelihood tree based on representative members actinobacterial classes and available full-length 16S rDNA sequence data from OPB41	90
Figure 4.2	Alignment of conserved signature indels in dUTP nucleotidohydrolase protein sequences	92
Figure 4.3	Alignment of conserved signature indels in adenylate kinase protein sequences.....	92
Figure 4.4	Maximum likelihood tree inferred from 128 single-copy proteins genomes in the Terrabacteria superphyla.....	94
Figure 4.5	Comparative analysis of protein clusters within all known Actinobacteria genomes.....	96
Figure 4.6	The number of coding sequences (CDS) in each COG categories relative to total number CDS in each set of protein clusters.....	101
Figure 4.7	Alignment of predicted ladderance biosynthetic gene clusters.....	103
Figure S4.1	Two-dimensional scaling of oligonucleotide signatures of assembled metagenomic sequence fragments from sediment in the Canterbury Basin off the coast of New Zealand	114

List of Attachments

Supplementary Data Table 2.1	Microsoft Excel worksheet containing all protein clusters with annotations	Supplemental_Data_Sheet_1.xlsx
Supplementary Data Table 3.1	Microsoft Excel worksheet containing all protein annotation and transcript recruitment values	Supplemental_Data_Sheet_2.xlsx

Chapter I: Introduction

Marine sediments contain an unimaginably large number of microbes, with current estimates at 2.9×10^{29} cells worldwide (1). Moreover, marine sediments are home to the highest proportion of uncultured genera (91-92%) among biomes frequently targeted with 16S RNA marker gene analysis (2), in which fragments of a universally conserved gene are amplified from the environment, ligated into a selectable plasmid, transformed into a cultivable strain of *Escherichia coli*, and sequenced after being extracted. Obtaining marine sediment organisms involves anything from using small push cores to commissioning massive ocean drilling platforms (Fig. 1.1, Fig. 1.2). Innovative “omics” techniques described below are then applied in order to unearth “intraterrestrials” or “microbial dark matter”, so named due to their ubiquity and uncultured status (3).

While single-cell genomics can be accomplished using various techniques, fluorescence activated cell sorting (FACS) is most commonly used in the field of environmental microbiology and at service facilities, such as the Single Cell Genomics Center at Bigelow Laboratory for Ocean Sciences (4). FACS techniques have been used in immunological studies for decades to separate out fluorescently tagged cell lines and, in recent years, to investigate mutations in individual cancer cells (5, 6). In FACS-based single cell research, cells are physically separated and forced through a narrow stream of buffer where they can be sorted based on natural or dye-associated fluorescence and subsequently lysed open. The genomes are then amplified and sequenced using next-generation sequencing platforms (Figure 1.3). Lloyd et al were the first to apply single-cell sequencing to marine sediment samples, optimizing cell separation from the sediment matrix, sorting and sequencing to obtain genomes from novel archaeal phyla (7).

Another technique developed at the same time to obtain the genomes of single organisms is metagenomic binning (8) (Figure 1.3). In this technique, DNA is extracted directly from the environmental sample and sequenced before “reconstructing” coherent genomes of single organisms using sequence intrinsic signatures such as tetramer frequencies and/or abundance-based techniques (9, 10). With techniques based on sequence compositions, researchers take advantage of the codon bias and the high coding density of bacterial and archaeal genomes to determine which sequence fragments may have originated from a similar group of organisms. A genomic fragment contains a matrix of 4 nucleotides “words.” The matrices found in other genomic fragments from the same or closely related organisms are made up of a similar set of 4



Figure 1.1: Karen G. Lloyd, Richard K. Kevorkian, and Jordan T. Bird pictured at the White Oak River estuary in Stella, NC. Richard (pictured middle) is holding the first push cores that the Lloyd Lab retrieved from “Station H” (37).



Figure 1.2: The R/V Greatship Manisha is the mission specific retrofitted drilling platform that was used to collect rock and sediment samples during IODP Expedition 347.

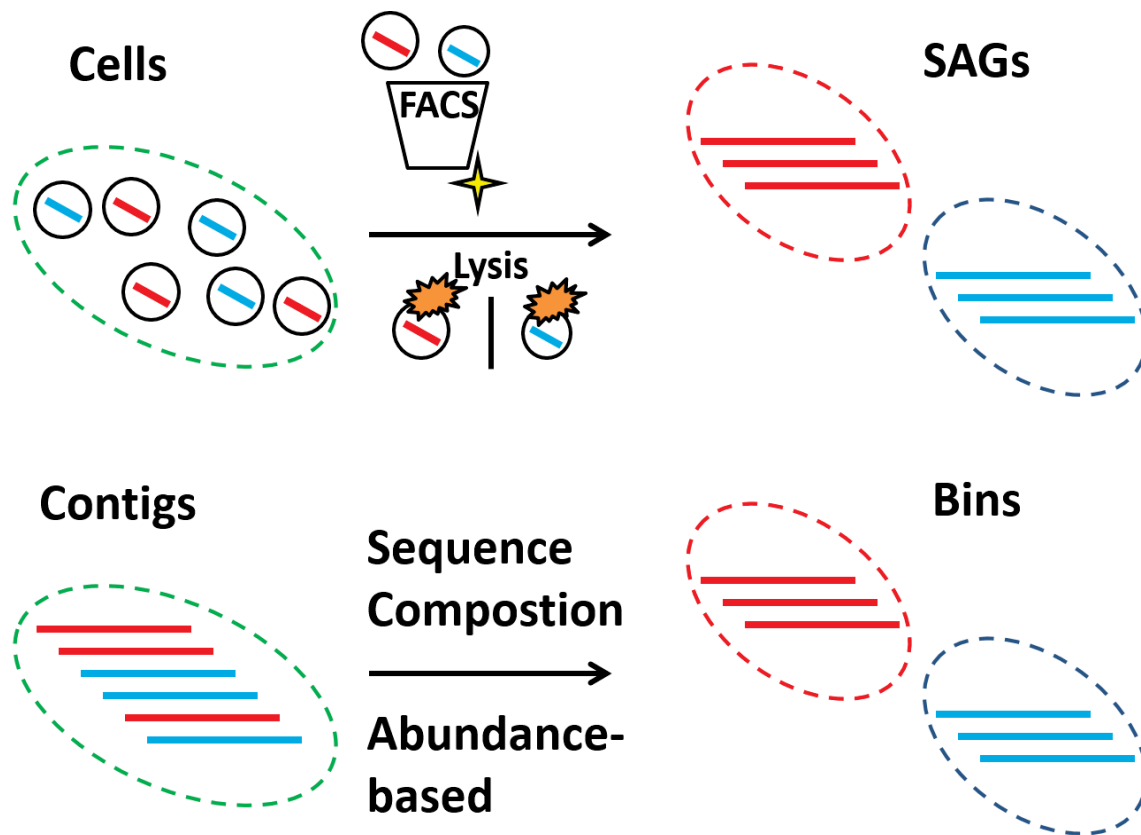


Figure 1.3: Schematic of single-cell genomics and metagenomic reconstruction techniques.

nucleotide words. Abundance-based techniques take advantage of the variable abundance of microbial species across separate metagenomic samples to tease apart assembled sequences with similar n-mer frequencies. Abundance-based techniques take advantage of the variable abundance of microbial species across separate metagenomic samples to tease apart assembled sequences with similar n-mer frequencies based on consistent abundance patterns across metagenomes.

The advantage of metagenomic binning is that it generates a much larger number of genomes from samples in the environment than single-cell genomics. The key disadvantage is that it is often impossible to resolve closely related and/or similarly abundant genomes into coherent genomes of a single organism. Additionally, 16S rRNA genes, which have defined microbial phylogeny across the domains of life, often don't bin with their corresponding genomic information because of the highly conserved nature of the gene. In some cases, single-cell genomics offers researchers a more precise tool for targeting specific taxa or functional metabolisms within a given environment (11–13).

Since the advent of metagenomics and single-cell genomic applications, the “Tree of Life” has grown by leaps and bounds (14). Over this time of expansion, comparative genomics has served as a tool to understand the evolutionary origins of the life and sparked intense scientific discussion (15). The group formerly known as Eocytes and then Crenarchaeota has exploded into an ever-growing list of dozens of suspected phylum-level taxa (14). The Eocyte hypothesis suggests that the domain Archaea is not monophyletic, but that one particular archaeal branch gave rise to present day Eukaryotes as shown by the similarities between Eocyta and Eukarya with regard to ribosomal structure and other “information processing” cellular machinery (15). Researchers analyzing genomes obtained from metagenomic binning and single-cell sequencing techniques have found new support of the Eocyte hypothesis (16–18). First proposed in 1984, the Eocyte hypothesis, which challenged Carl Woese and George Fox’s landmark paper proposing three domains of life, now has support from phylogenetic trees based on conserved proteins, as well as, the discovery of cytoskeletal machinery key to the development of invaginated cell membranes within the closest archaeal relatives of Eukarya (16–18, 18–27).

Besides finding the potential close relatives to the Last Eukaryotic Common Ancestors (LECA), metagenomics and single-cell genomics have allowed for many other discoveries in

phylogenomics and functional metagenomics. Researchers assembled genomes in samples from terrestrial aquifer fluids containing a radiation of microbial taxa, known simply as Candidate Phyla Radiation (CPR), that makes up now 15% of the phyla in the domain Bacteria (28). Similar studies have defined the first genomes from dozens of new bacterial and archaeal phyla (29–32). Each of these assembled genomes contains hundreds to thousands of genes. As the databases of genomic information continue to fill up with genes from these novel, uncultured clades it is important to ask questions about the physiology of these microbes, their specific geochemical context, and their shared evolutionary histories.

In part, this dissertation provides genomic analysis with environmental context for some of the novel phyla most commonly found in the marine sediments of the White Oak River estuary. The White Oak River estuary is an anoxic sediment environment where sediment biogeochemistry has been measured in scientific studies for the last 39 years (33–41). Many of the drivers of geochemical variability at this site are well understood, so the White Oak River estuary is a prime source of samples for microbiological and sequencing-based studies that seek to connect dynamic geochemical changes to the biological drivers of the organic matter degradation and anaerobic respiration (29, 41–47). Methane production and anaerobic methane oxidation are key biological metabolisms of the downstream portions of the estuary which are influenced by high sulfate overlying waters (41). While these metabolisms are clearly drivers of the biogeochemical change, they only support a small portion of the microbial diversity found in the sediments. The metabolisms of the majority of microbial taxa in these sediments are largely unknown, as no close relatives have been cultured. To this end, I have endeavored to use single-cell sequencing techniques to investigate the potential metabolisms of uncultivated archaeal taxa in this well-studied coastal sediment environment.

Within White Oak River Estuary sediment, abundances of archaea match those of bacteria (42). Archaea within the phylum Euryarchaeota are the only organisms known to both produce and consume methane anaerobically. Uncultivated archaeal taxa previously identified in White Oak River clone libraries include abundant members of the Bathyarchaeota, Marine Group 1, Terrestrial Hot Springs Group, Marine Benthic Group B, ANME-1, Thermoplasmatales, and Deep Sea Euryarchaeota Group, as well as, other sequences clustering within the large phyla of Crenarchaeota and Euryarchaeota (42). Evidence from the mRNA transcript qPCR of ANME-1

methyl co-enzyme B reductase (*mcr*) suggests that members of this branch of Euryarchaeota, previously associated only with methane consumption, also produce methane (41). Recent evidence from another marine sediment site detected an *mcr* complex within Bathyarchaeotal genome reconstructions from another marine sediment site (48), but this gene has not been detected among several Bathyarchaeota genome reconstructions recovered from White Oak River Estuary sediments (47). Questions still remain about the potential metabolism of the other uncultured archaea taxa within the White River Estuary.

In 2013, IODP Expedition 347 set out to collect deep sediment cores from the Baltic Sea. This site was selected because it records the relatively recent sedimentary history across the latest glacial maxima with the oldest samples for biological investigation coming from 44,000-year-old sediments (49) (Fig. 1.4). Glacial coverage of the Baltic Sea temporarily shifted the sedimentary regimes from a coastal brackish/marine system to a lacustrine depositional environment. In contrast to the White Oak River Estuary, bacteria in Baltic Sea sediments outnumber archaea by up to two orders of magnitude (50). Included in the post expedition scientific report was the stated objective “to determine the diversity and activity of subsurface microbial communities through metagenomic, metatranscriptomic, and single-cell genomic analyses” (49). In-depth metagenomic and metatranscriptomic analyses of the Baltic Sea sediments were performed by our collaborators on this project (51, 52). Genomes assembled from the recovered metagenomes did not provide long enough contiguous sequences to make metagenomic binning techniques efficacious (52). This work instead details environmental transcriptomic analysis similar to methods used in other studies in which transcripts from the environment are mapped back to genomic contigs obtained through a separate sequencing effort (53–56). Additionally, untargeted metabolomic techniques similar to those used in other laboratory-based and environmental samples were used for the first time to investigate the small molecule pools available to Baltic Sea sediment microorganism (55, 57). These were combined with direct enzyme assays of key carbon-degrading enzymes predicted in the SAGs.

Microbial persistence has been described as a bet-hedging strategy, where a portion of a microbial population enters a dormancy of low to zero growth (58, 59). Mechanisms such as

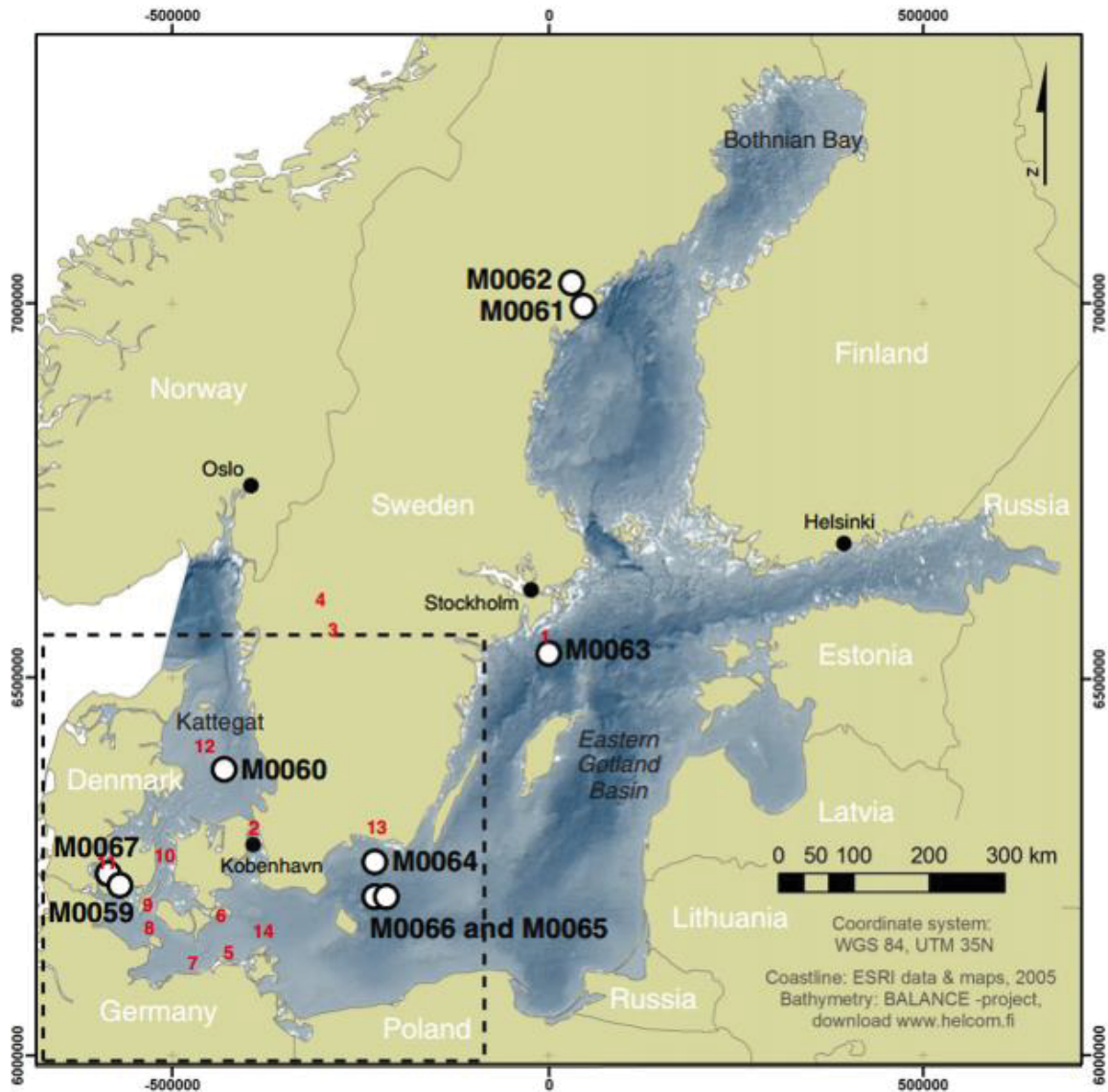


Figure 1.4: Map of sample sites from IODP Expedition 347: Baltic Sea Paleoenvironments. Drilling sites discussed in this work include M0059, M0060, and M0063. Red numbers indicate relevant geological formations: 1 = Landsort Deep, 2 = Öresund Strait, 3 = South Central Sweden, 4 = Lake Vänern, 5 = Darss, 6 = Møn, 7 = Mecklenburger Bay, 8 = Fehmar Belt, 9 = Langeland, 10 = Great Bält, 11 = Little Bält, 12 = Anholt, 13 = Blekinge archipelago, 14 = Kriegers Flak. This figure is a reprint of one that first appeared in the preliminary cruise report (49).

toxin-antitoxin genomic features and stringent response transcriptional effectors have been proposed to play a role in persister cell development (58, 59). Other researchers have attempted to explain greater than expected diversity in limited-energy environments and persistence strategies such as avoiding decay through inactivity by postulating a seed-bank theory (60, 61). Modeling studies have suggest that the relative size of inactive populations scale dramatically with increased residence time (62).

Microbial persistence has been described as a bet-hedging strategy, where a portion of a microbial population enters a dormancy of low to zero growth (58, 59). Mechanisms such as toxin-antitoxin genomic features and stringent response transcriptional effectors have been proposed to play a role in persister cell development (58, 59). Other researchers have attempted to explain greater than expected diversity in limited-energy environments by postulating a seed-bank theory (60, 61). Seed banks similarly described subpopulations of persister cells avoiding decay through inactivity. Modeling studies suggest that the relative size of inactive populations scale dramatically with increased residence time (62).

Marine sediments are characterized by particularly long residence times with fine-grained clay-rich sediments like those found in the Baltic Sea described as relatively impermeable (63). Marine sediments have a sedimentation rate ranging from 0.0001 cm to 10 cm per year (64, 65), while biological carbon turnover times are only slightly above the biochemical limit set by the racemization rate of amino acids (66). Marine sediments like the Baltic Sea sediment are natural laboratories that allow us to investigate the key metabolisms of microbes that may be able to persist in these low-energy environments for thousands of years according to current estimates of population doubling time (66, 67).

The candidate phylum Atribacteria is prevalent in deep sea sediment communities appearing in 66% of 16S rRNA clone libraries greater than 2 meters below seafloor (mbsf) according to one review of 205 16S rRNA gene libraries from subsurface marine sediments (68). Atribacteria also is among a small number of dominant phyla in these datasets along with Chloroflexi, Gammaproteobacteria, and Planctomycetes. Other bacteria phyla averaging greater than 1% of the sampled bacterial communities were Alpha-, Beta-, Delta- and Epsilonproteobacteria, Bacteroidetes, Firmicutes, Actinobacteria, Spirochaetes, Aminicenantes, Microgenomates and Aerophobetes. Predicted metabolisms for Atribacteria include both

propionate and sugar fermentation (69–71). Other recently named groups with no cultured members including Aminicenantes, Microgenomates, and Aerophobetes were also predicted to be capable of fermentation but not respiration (69, 72). In an analysis that combined 16S rRNA clone libraries and terminal-restriction fragment length polymorphism analyses across 8,000 years of marine sediment deposition at the bottom of the Baltic Sea, most of the partial 16S rRNA genes were from members of the Atribacteria phyla (73).

In the final chapter of this dissertation, I compare the currently available genomic and biogeographical data regarding a novel taxon within the Terrabacteria. Terrabacteria is a superphylum that includes five phyla with cultured members (74). In one widely considered evolutionary hypothesis, life was confined to hot, organic molecule-rich oceans and aquatic environments throughout most of Earth's early microbial history (75). While Terrabacteria includes some host-associated microorganisms including *Mycobacteria tuberculosis* and *Streptococcus pyogenes*, the vast majority of genera within the Terrabacteria superphylum contain members that possess key adaptations to land colonization such as pigmentation and resistance to desiccation (74). Comparative genomics has been used as tool to better understand the origins of microbial life on land through the Terrabacteria (76–78). Understanding phylogenetic placement and functional diversity within a deeply branching taxa of Terrabacteria informs the current understanding of the ancestral microorganisms that took of life from the sea to land.

List of References

1. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A* 109:16213–16216.
2. Lloyd KG, Bird JT, Deas E, Kevorkian R, Ladau J, Noordhoek T, Roy T, Yin J, Crosby L. 2016. The nature and environmental distribution of Earth's uncultured microbiome. *Prep*.
3. Valentine DL. 2013. Microbiology: Intraterrestrial lifestyles. *Nature* 496:176–177.
4. Stepanauskas R. 2012. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* 15:613–620.
5. Macaulay IC, Voet T. 2014. Single Cell Genomics: Advances and Future Perspectives. *PLOS Genet* 10:e1004126.
6. Saadatpour A, Lai S, Guo G, Yuan G-C. 2015. Single-Cell Analysis in Cancer Genomics. *Trends Genet* 31:576–586.
7. Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD, Stepanauskas R, Richter M, Kleindienst S, Lenk S, Schramm A, Jørgensen BB. 2013. Predominant archaea in marine sediments degrade detrital proteins. *Nature* 496:215–218.
8. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10:doi: 10.1186/gb-2009-10-8-r85.
9. Sangwan N, Xia F, Gilbert JA. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4:8.
10. Sedlar K, Kupkova K, Provaznik I. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* 15:48–55.
11. Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, Holland T, Cotton D, Hauser L, Keller M. 2007. Targeted Access to the Genomes of Low-Abundance Organisms in Complex Microbial Communities. *Appl Environ Microbiol* 73:3205–3214.

12. Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, Xie G, Poulton NJ, Gomez ML, Masland DED, Thompson B, Bellows WK, Ziervogel K, Lo C-C, Ahmed S, Gleasner CD, Detter CJ, Stepanauskas R. 2012. Capturing Single Cell Genomes of Active Polysaccharide Degradors: An Unexpected Contribution of Verrucomicrobia. *PLOS ONE* 7:e35314.
13. Campbell AG, Campbell JH, Schwientek P, Woyke T, Sczyrba A, Allman S, Beall CJ, Griffen A, Leys E, Podar M. 2013. Multiple Single-Cell Genomes Provide Insight into Functions of Uncultured Deltaproteobacteria in the Human Oral Cavity. *PLOS ONE* 8:e59361.
14. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nat Microbiol* 1:doi: 10.1038/nmicrobiol.2016.48.
15. Lake JA. 2015. Eukaryotic origins. *Phil Trans R Soc B* 370:20140321.
16. Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504:231–236.
17. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521:173–179.
18. Guy L, Ettema TJG. 2011. The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol* 19:580–587.
19. Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci* 74:5088–5090.
20. Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc Natl Acad Sci* 81:3786–3790.
21. Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* 331:184–186.

22. Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci* 87:4576–4579.
23. Rivera MC, Lake JA. 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257:74–76.
24. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci* 105:20356–20361.
25. Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV. 2008. The deep archaeal roots of eukaryotes., *The Deep Archaeal Roots of Eukaryotes. Mol Biol Evol* Mol Biol Evol 25, 25:1619, 1619–1630.
26. Kelly S, Wickstead B, Gull K. 2011. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc Biol Sci* 278:1009–1018.
27. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541:353–358.
28. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523:208–211.
29. Baker BJ, Lazar CS, Teske AP, Dick GJ. 2015. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* 3:14.
30. Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, Banfield JF. 2016. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations. *Environ Microbiol* ‘Accepted Article,’ doi: 10.1111/1462-2920.13362.
31. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF.

2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 7.
32. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 1.
33. Martens CS, Goldhaber MB. 1978. Early diagenesis in transitional sedimentary environments of the White Oak River Estuary, North Carolina. *Limnol Oceanogr* 23:428–441.
34. Chanton JP, Martens CS, Kipphut GW. 1983. Lead-210 sediment geochronology in a changing coastal environment. *Geochim Cosmochim Acta* 47:1791–1804.
35. Chanton JP, Martens CS. 1988. Seasonal variations in ebullitive flux and carbon isotopic composition of methane in a tidal freshwater estuary. *Glob Biogeochem Cycles* 2:289–298.
36. Chanton JP, Martens CS, Kelley CA. 1989. Gas transport from methane-saturated, tidal freshwater and wetland sediments. *Limnol Oceanogr* 34:807–819.
37. Kelley CA, Martens CS, Chanton JP. 1990. Variations in sedimentary carbon remineralization rates in the White Oak River estuary, North Carolina. *Limnol Oceanogr* 35:372–383.
38. Kelley CA, Martens CS, Ussler W. 1995. Methane dynamics across a tidally flooded riverbank margin. *Limnol Oceanogr* 40:1112–1129.
39. Hoehler TM, Alperin MJ, Albert DB, Martens CS. 1994. Field and laboratory studies of methane oxidation in an anoxic marine sediment: Evidence for a methanogen-sulfate reducer consortium. *Glob Biogeochem Cycles* 8:451–463.
40. Kogan I, Paull CK. 2005. Coastal seismic wipe-outs: Distribution controlled by pore water salinity. *Mar Geol* 217:161–175.
41. Lloyd KG, Alperin MJ, Teske A. 2011. Environmental evidence for net methane production and oxidation in putative ANaerobic METHanotrophic (ANME) archaea. *Environ Microbiol* 13:2548–2564.

42. Kubo K, Lloyd KG, Biddle JF, Amann R, Teske A, Knittel K. 2012. Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J* 6:1949–1965.
43. Lazar CS, Baker BJ, Seitz K, Hyde AS, Dick GJ, Hinrichs K-U, Teske AP. 2015. Genomic evidence for distinct carbon substrate preferences and ecological niches of Bathyarchaeota in estuarine sediments. *Environ Microbiol* 18:1200–1211.
44. Meador TB, Bowles M, Lazar CS, Zhu C, Teske A, Hinrichs K-U. 2015. The archaeal lipidome in estuarine sediment dominated by members of the Miscellaneous Crenarchaeotal Group. *Environ Microbiol* 17:2441–2458.
45. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J* [published online ahead of print]:doi: 10.1038/ismej.2015.233.
46. Baker BJ, Saw JH, Lind AE, Lazar CS, Hinrichs K-U, Teske AP, Ettema TJG. 2016. Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nat Microbiol* 1:16002.
47. Lazar CS, Baker BJ, Seitz KW, Teske AP. 2017. Genomic reconstruction of multiple lineages of uncultured benthic archaea suggests distinct biogeochemical roles and ecological niches. *ISME J* 11:1058.
48. He Y, Li M, Vengadesh perumal N, Feng X, Fang J, Xie J, M. Sievert S, Wang F. 2016. Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat Microbiol* 1:16035.
49. Andrén T, Jørgensen B, Cotterill C, Green S. 2015. IODP expedition 347: Baltic Sea basin paleoenvironment and biosphere. *Sci Drill* 20:1–12.
50. Buongiorno J, Turner S, Webster G, Asai M, Shumaker AK, Roy T, Weightman A, Schippers A, Lloyd KG. 2017. Interlaboratory quantification of Bacteria and Archaea in deeply buried sediments of the Baltic Sea (IODP Expedition 347). *FEMS Microbiol Ecol* 93.

51. Zinke LA, Mullis MM, Bird JT, Marshall IPG, Jørgensen BB, Lloyd KG, Amend JP, Reese BK. 2017. Thriving or Surviving? Evaluating active microbial guilds in Baltic Sea sediment. *Environ Microbiol Rep* [Accepted].
52. Marshall IPG, Karst SM, Nielsen PH, Jørgensen BB. 2017. Metagenomes from deep Baltic Sea sediments reveal how past and present environmental conditions determine microbial community composition. *Mar Genomics*.
53. Baker BJ, Lesniewski RA, Dick GJ. 2012. Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6:2269–2279.
54. Lesniewski RA, Jain S, Anantharaman K, Schloss PD, Dick GJ. 2012. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J* 6:2257–2268.
55. Steffen MM, Dearth SP, Dill BD, Li Z, Larsen KM, Campagna SR, Wilhelm SW. 2014. Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in *Microcystis*. *ISME J* 8:2080–2092.
56. Voorhies AA, Eisenlord SD, Marcus DN, Duhaime MB, Biddanda BA, Cavalcoli JD, Dick GJ. 2015. Ecological and genetic interactions between cyanobacteria and viruses in a low-oxygen mat community inferred through metagenomics and metatranscriptomics. *Environ Microbiol*.
57. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD. 2006. Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci* 103:19302–19307.
58. Maisonneuve E, Gerdes K. 2014. Molecular Mechanisms Underlying Bacterial Persisters. *Cell* 157:539–548.
59. Harms A, Maisonneuve E, Gerdes K. 2016. Mechanisms of bacterial persistence during stress and antibiotic exposure. *Science* 354:aaf4268.
60. E Jones S, Lennon J. 2010. Jones SE, Lennon JT.. Dormancy contributes to the maintenance of microbial diversity. *Proc Nat Acad Sci USA* 107: 5881-5886. *Proc Natl Acad Sci U S A* 107:5881–6.

61. Lennon J, E Jones S. 2011. Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* 9:119–30.
62. Locey KJ, Lennon JT. 2017. A residence-time framework for biodiversity. e2727v2. *PeerJ Preprints*.
63. Moore DG, Keller GH. Marine sediments, geotechnical properties Marine sediments, geotechnical properties. *Encycl Earth Sci Ser Appl Geol* 343–350.
64. Elliott EA, McKee BA, Rodriguez AB. 2015. The utility of estuarine settling basins for constructing multi-decadal, high-resolution records of sedimentation. *Estuar Coast Shelf Sci* 164:105–114.
65. Schofield MM, Jain S, Porat D, Dick GJ, Sherman DH. 2015. Identification and analysis of the bacterial endosymbiont specialized for production of the chemotherapeutic natural product ET-743. *Environ Microbiol* 17:3964–3975.
66. Hoehler TM, Jørgensen BB. 2013. Microbial life under extreme energy limitation. *Nat Rev Microbiol* 11:83–94.
67. Braun S, Mhatre SS, Jaussi M, Røy H, Kjeldsen KU, Pearce C, Seidenkrantz M-S, Jørgensen BB, Lomstein BA. 2017. Microbial turnover times in the deep seabed studied by amino acid racemization modelling. *Sci Rep* 7:5680.
68. Parkes RJ, Cragg B, Roussel E, Webster G, Weightman A, Sass H. 2014. A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere:geosphere interactions. *Mar Geol* 352:409–425.
69. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.
70. Carr SA, Orcutt BN, Mandernack KW, Spear JR. 2015. Abundant Atribacteria in deep marine sediment from the Adélie Basin, Antarctica. *Front Microbiol* 6: Article 872.
71. Nobu MK, Dodsworth JA, Murugapiran SK, Rinke C, Gies EA, Webster G, Schwientek P, Kille P, Parkes RJ, Sass H, Jørgensen BB, Weightman AJ, Liu W-T, Hallam SJ,

- Tsiamis G, Woyke T, Hedlund BP. 2015. Phylogeny and physiology of candidate phylum ‘Atribacteria’ (OP9/JS1) inferred from cultivation-independent genomics. *ISME J*.
72. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF. 2012. Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* 337:1661–1665.
 73. Lyra C, Sinkko H, Rantanen M, Paulin L, Kotilainen A. 2013. Sediment Bacterial Communities Reflect the History of a Sea Basin. *PLOS ONE* 8:e54326.
 74. Battistuzzi FU, Hedges SB. 2009. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. *Mol Biol Evol* 26:335–343.
 75. Baross JA, Hoffman SE. 1985. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Orig Life Evol Biosph* 15:327–345.
 76. Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 4:44.
 77. Wu H, Fang Y, Yu J, Zhang Z. 2014. The quest for a unified view of bacterial land colonization. *ISME J* 8:1358–1369.
 78. Ghylis TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, Sczyrba A, Stepanauskas R, Grossart H-P, Woyke T, Warnecke F, Malmstrom R, Bertilsson S, McMahon KD. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *ISME J*.

Chapter II: Culture independent genomic comparisons reveal environmental adaptations for Altiarchaeales

This chapter was previously published as a peer-reviewed article in *Frontiers in Microbiology* authored by Jordan T. Bird, Brett J. Baker, Alexander J. Probst, Mircea Podar, and Karen G. Lloyd.

Jordan T. Bird and Karen G. Lloyd designed the study. All analyses were completed by Jordan T. Bird, Karen G. Lloyd, Mircea Podar, and Brett J. Baker. Other genomic data for comparative analyses was provided by Brett J. Baker and Alexander J. Probst. All authors contributed to the discussion of the results. Manuscript was written by Jordan T. Bird with inputs from Karen G. Lloyd, Brett Baker, Mircea Podar, and Alexander J Probst.

Abstract

The recently proposed candidatus order Altiarchaeales remains an uncultured archaeal lineage composed of genetically diverse, globally distributed organisms frequently observed in anoxic subsurface environments. In spite of 15 years of studies on the psychrophilic biofilm-producing *Candidatus* Altiarchaeum hamiconexum and its close relatives, very little is known about the phylogenetic and functional diversity of the widespread free-living marine members of this taxon. From methanogenic sediments in the White Oak River Estuary, NC, USA, I sequenced a single amplified genome (SAG), WOR_SM1_SCG, and used it to recover and refine two high-quality metagenome-assembled genomes (MAGs), WOR_SM1_79 and WOR_SM1_86-2, sampled from the same site on a different date. Our analysis of these three genomic reconstructions indicates that they are part of a monophyletic group which includes three previously published MAGs from terrestrial springs and a SAG from Sakinaw Lake in a group previously designated as pMC2A384. For each, a synapomorphic mutation in the Altiarchaeales tRNA synthetase β subunit, *pheT*, causes the protein to be encoded as two subunits at non-adjacent loci. Consistent with the terrestrial spring clades, our estuarine genomes contain a near-complete autotrophic metabolism: H₂ or CO as potential electron donors, a reductive acetyl-CoA pathway for carbon fixation, and methylotroph-like NADP(H)-dependent dehydrogenase. Phylogenies generated from 16S rRNA genes and concatenated conserved proteins identify two distinct sub-clades of Altiarchaeales: Alti-1 populated by organisms from actively flowing springs and Alti-2, which was more widespread, more diverse, and not associated with visible mats. The core Alti-1 genome suggest Alti-1 has adapted to the stream environment through the development of lipopolysaccharide production capacity and extracellular hami structures. The core Alti-2 genome suggest members of this clade are free-living with distinct mechanisms for energy maintenance, motility, osmoregulation, and sulfur redox reactions. These data suggest that the hamus structures found in *Candidatus* Altiarchaeum hamiconexum are not present outside of stream-adapted Altiarchaeales. Homologs to a Na⁺ transporter and membrane bound coenzyme A disulfide reductase that were unique to the brackish sediment Alti-2 genomes could indicate adaptations to the estuarine, sulfur-rich environment.

Introduction

An uncultivated group of environmental archaea, originally called the SM1 but recently given the name Altiarchaeales, was first described as a nearly monoclonal biofilm in a cold terrestrial sulfidic spring in Regensburg, Germany ([Rudolph et al., 2001](#); [Moissl et al., 2002](#)). The SM1 was originally described as having unique appendages called hami that work like grappling hooks to maintain their position in the flowing springs ([Moissl et al., 2002, 2003, 2005](#)). Altiarchaeales represent one example of biofilm-forming archaea, along with *Sulfolobus* sp. in hot springs ([Brock et al., 1972](#); [Zillig et al., 1980](#); [Suzuki et al., 2002](#)), and uncultured ANME archaea in euxinic basins ([Michaelis et al., 2002](#)). Their hami structures have no analog in other microbes and might have technological importance due to their intricate nano-sized structure ([Perras et al., 2014](#)). Additionally, Altiarchaeales appear to be one of the few examples of archaea with a double cell membrane ([Probst et al., 2014](#); [Probst and Moissl-Eichinger, 2015](#)). Furthermore, the Altiarchaeales appear to belong to the phylum Euryarchaeota, which contains most of the industrially and environmentally important archaeal cultures: halophilic phototrophs, sulfate reducers, iron cycling extremophiles, and all cultured methanogens. However, little is known about the functional diversity and evolutionary history of the Altiarchaeales. 16S rRNA gene diversity surveys have indicated that Altiarchaeales are a globally distributed group with a wide preference for anoxic environments such as lake sediments, sulfidic aquifers, geothermal springs, deep sea sediments, mud volcanoes, and hydrothermal vents as well as industrial settings and drilled wells ([Probst et al., 2014](#)) (Figure 2.1).

Despite the cosmopolitan nature of the Altiarchaeales, these organisms have never been isolated in pure culture, and previously published MAGs have only been obtained from terrestrial cold springs. A metagenome amplified from samples of visible mats floating in a sulfidic spring (Muehlbacher Schwefelquelle, Germany) enabled the assembly of the *Candidatus* Altiarchaeum hamiconexum genome, MSI_SM1 ([Probst et al., 2014](#)). MSI_SM1 contains putative genes for the hami as well as conserved evolutionary marker genes that distinguish it as a new order within the Euryarchaeota ([Probst et al., 2014](#)). *Candidatus* Altiarchaeum hamiconexum is naturally enriched in sulfidic springs and has been hypothesized to play a role in sulfur cycling ([Moissl et al., 2002](#)). However, MSI_SM1 contains no genetic evidence for the use of sulfur-containing compounds in respiration. A genome from a less abundant Altiarchaeales, IMC4_SM1,

reconstructed from the same sample, and another genome reconstructed from subsurface water filtrates at the Crystal Geyser (USA) spring, CG_SM1, is closely related to *Candidatus* Altiarchaeum hamiconexum ([Probst et al., 2014](#)). In both cases, these microbes are dominant members of their microbial communities. In-depth genomic analyses of MSI_SM1 and CG_SM1 have suggested that the Altiarchaeales are autotrophic, utilizing a modified version of the archaeal reductive acetyl-CoA (Wood–Ljungdahl) pathway. Further support for autotrophy comes from the ^{13}C -depleted isotope content of the lipid archaeol found at the German site ([Probst et al., 2014](#)). With >98% identical 16S rRNA genes, MSI_SM1 and CG_SM1 share close evolutionary histories. Furthermore, all three MAGs were taken from similar terrestrial cold spring environments. Therefore, in order to describe the functional diversity and evolutionary radiation of the order Altiarchaeales, it is important to expand our genomic comparison to include distantly related members obtained from different environments.

I obtained genomic reconstructions from brackish sediments in the White Oak River Estuary (WOR), NC, USA. These sediments have a stable redox gradient with microbially mediated sulfate reduction via organic matter oxidation, followed by methane oxidation at the sulfate methane transition zone (SMTZ), and finally methanogenesis where sulfate is depleted ([Martens and Goldhaber, 1978](#); [Kelley et al., 1990](#); [Lloyd et al., 2011](#)). The microbial community is similar to those found in marine sediments worldwide ([Lloyd et al., 2011](#); [Kubo et al., 2012](#)).

Similar to the German spring environments and Crystal Geyser, the WOR sediment environment is sulfidic ([Rudolph et al., 2004](#); [Probst et al., 2016](#)). However, microbial communities deep in the WOR sediment do not experience the active flow regimes of those microbial communities in the German spring and Crystal Geyser. The WOR sediments have much more stable flow regimes and redox gradients that shift on longer seasonal timescales ([Lloyd et al., 2011](#)). In WOR sediments, microorganisms demonstrate a clear effect on the geochemical environment as vertical profiles often reveal hierarchical zonation based on available highest energy electron acceptors ([Kelley et al., 1990](#)). Furthermore, deep in WOR sediments, the temperature varies seasonally, not daily, and rain events have minimal impact. Although a 16S rRNA gene from Altiarchaeales has been found in WOR, they do not make visible biomass and are likely less dominant than the Altiarchaeales found in the German sulfidic

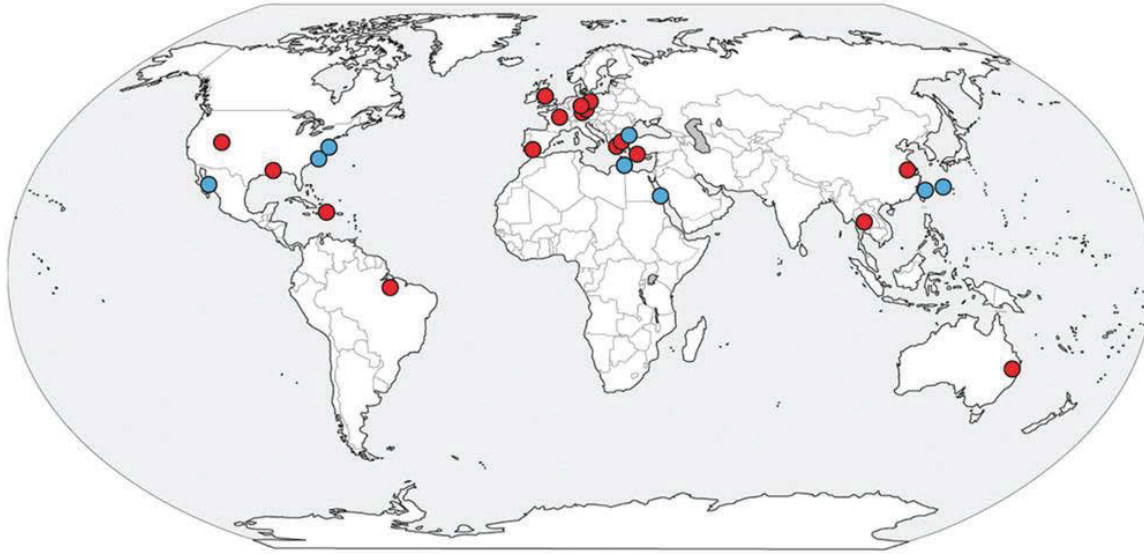


Figure 2.1: Global distribution of Altiarchaeales 16S rRNA gene sequences present in the NCBI database.

springs relative to the total biomass of the respective environmental niches ([Rudolph et al., 2004](#); [Kubo et al., 2012](#); [Probst et al., 2014, 2016](#)).

I obtained a single cell genome from a methane-rich depth (Figure 2.2) in the WOR and used it to recover two MAGs sampled from the same site on a different date. Dozens of other novel archaea and bacteria genomes have been derived from this dataset and are described elsewhere ([Baker et al., 2015](#); [Lazar et al., 2015](#); [Seitz et al., 2016](#)). By placing these new genomic reconstructions in the context of those previously recovered from sulfidic springs, I investigated the phylogeny of the Altiarchaeales and gained clues about how members of the Altiarchaeales show genetic adaptations for different environments.

Materials and Methods

Sampling

Plunger cores (1 m) of sediment were manually retrieved from 1.5 m of water in the WOR Station H (34° 44.490' N, 77° 07.44' W) in October 2010 and October 2012. Subsampling of the 2010 sample, and subsequent metagenomic sequencing were described previously ([Baker et al., 2015](#)). The 2012 core was sectioned within 30 h of retrieval into 3 cm intervals and samples were removed for methane and sulfate analyses, which were measured with the methods described in [Lloyd et al. \(2011\)](#). Sediments from the 72–75 cm section were used to fill up a 15 ml plastic tube and placed on water ice for transport to the University of Tennessee in Knoxville, Knoxville, TN, USA. After 48 h, 5 ml of anoxic artificial seawater (ASW) and ~1 g of sediment were added to a 10 ml glass serum vial under nitrogen gas. The serum vial was incubated between 20 and 27°C while shaking at 100 rpm for 48 h.

Cell Extraction

Cell extraction methods followed a previously published one ([Lloyd et al., 2013](#)), with the following minor modifications. The glass serum vial was shaken, the cap removed, and 5 ml of the colloidal mixture were transferred to a sterile 15 ml plastic tube before starting the cell extraction procedure. An additional 5 ml of anoxic ASW was added to the vial and transferred to the 15 ml tube to ensure complete sediment transfer. The tip of a Misonix Microson™

Ultrasonic Cell Disruptor was placed in an ice bath next to the plastic tube containing the sample and manually sonicated two times/s at 20% power to dislodge cells from their associated minerals. Next, the tube was vortexed for 10 s and the sediment was allowed to settle for 10 min.

Seven hundred and fifty microliter aliquots of the supernatant were gently laid on top of an equal volume of 60% Nycodenz solution in a sterile 2 ml microcentrifuge tube. The 2 ml tubes were then centrifuged at $11,617 \times g$ and 4°C for 1 h. The layer above the Nycodenz gradient was carefully removed and pooled into a sterile 15 ml plastic tube. Lastly, 1.5 ml aliquots were mixed with 375 μl of a 6% betaine 1x TE solution and immediately placed at -80°C .

Cell Sorting, Amplification, Screening, and Sequencing

A 1:1 mixture of extracted cells and sterile, 0.2 μm filtered, UV treated phosphate buffered saline (PBS) was incubated with 2 μl of 0.5 μM Syto 9 and 0.5 μM Syto 62 dye two 96 well plates were treated with UV for 30 min before 3 μl of TE were added to each well. The stained cells were gravity filtered through a 30 μm mesh and sorted with a BD Cytopeia Influx Flow Cytometer (Oak Ridge National Laboratory in Oak Ridge, TN, USA) in a class 1000 clean room. One cell was placed in each well based on sidescatter and 640 nm red fluorescence intensity. Sorted cells were lysed and DNA was amplified by multiple displacement amplification (MDA) as described previously (Beall et al., 2014). 16S rRNA genes were amplified from each well after a 1:150 dilution of the MDA products (SAGs) using primers A344f and A915r for Archaea (Stahl and Amann, 1991) and BAC-8F and BAC-1492R for bacteria (Teske et al., 2002). Products were visualized on an agarose gel. Six SAGs amplified with bacterial primers, six with archaeal. Samples amplifying with bacteria or both primer sets were not analyzed further. Samples that were positive for only archaea were confirmed by repeating the amplification. Only one SAG had replicable amplification when archaeal primers were used. After Sanger sequencing (UT Genomic Core) confirmed the SAG was archaeal, it was targeted for whole genome sequencing. The SAG DNA was quantified by UV absorbance, yielding 14 μg . DNA sequencing was performed on a TruSeq v2 type library with the Illumina MiSeq platform, 250 base pairs pair-end at the Hudson Alpha Genomics Center. The sample was treated with S1 nuclease prior to library preparation in order to cleave branched DNA structure formed by the MDA process ([Zhang et al., 2006](#)).

SAG Assembly and Annotation

The 19,275,108 sequences totaling over 48 million bp were quality filtered using the CLC Genomics Workbench 6.5.21 using the following criteria: >30 Phred score or trimmed on the ends until that criteria is met. Reads mapping at $>90\%$ length and $>98\%$ sequence identity to

Methane and Sulfate Concentrations Downcore

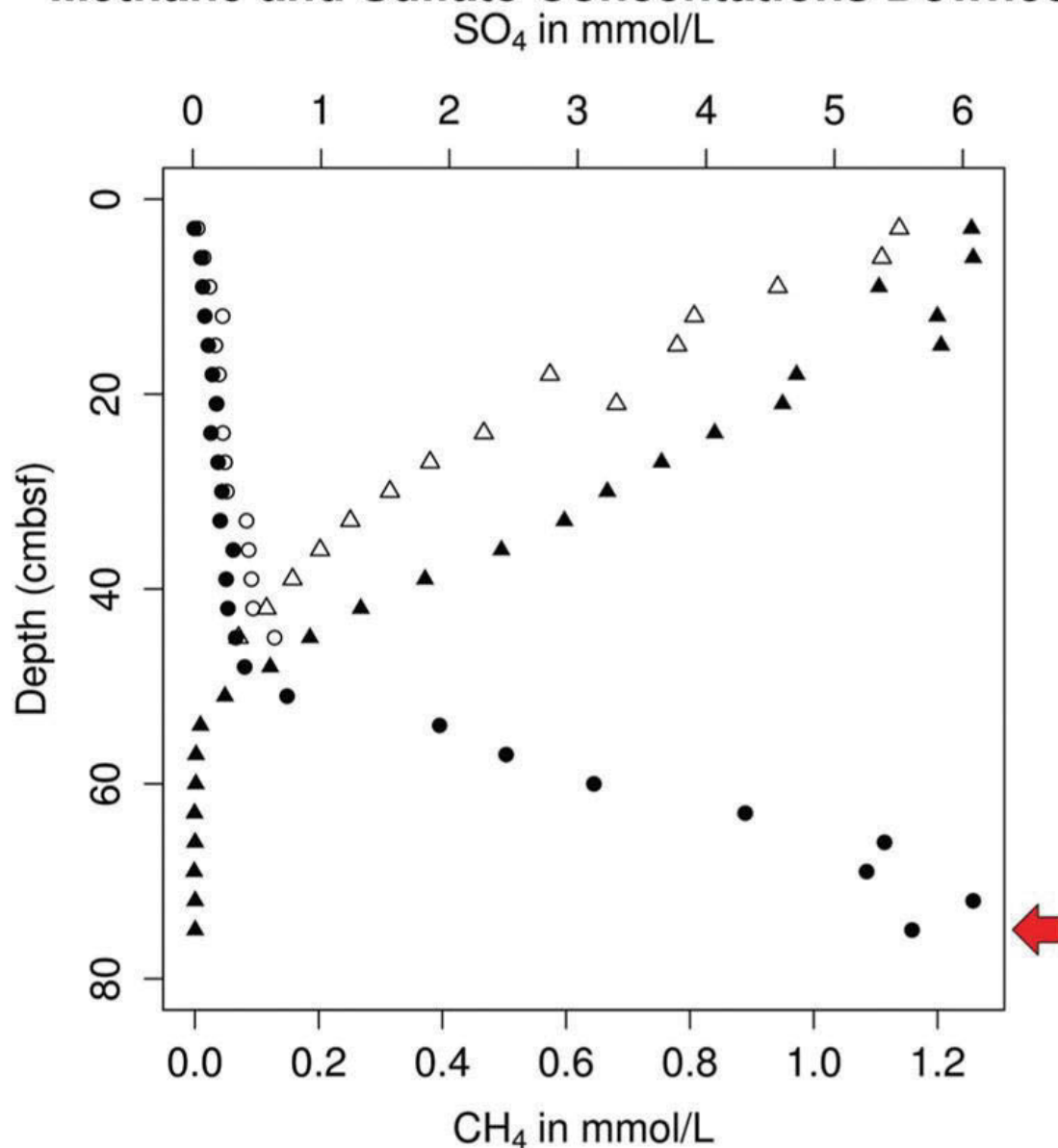


Figure 2.2: Aqueous porewater concentrations of sulfate (triangles) and methane (circles) in two replicate sediment cores (open/filled-in) taken from White Oak River estuary Station H in October 2012. The red arrow indicates the depth at which samples for single-cell sorting were taken.

common contaminants [i.e., *Homo sapiens*, *Escherichia coli* K12 (GC.1), *Saccharomyces cerevisiae* (GCA_000146045.2), Phi179 and Phi29] were removed along with duplicated sequences. 17,025,806 paired and 879,071 unpaired sequences between 20 and 251 bp remained. The remaining high quality paired and single reads were assembled using the SPAdes 3.0 assembler with the following parameters: `spades.py -m 62 -o <OUTPUT_DIR> -sc -12 <PAIRED_READS> -s <SINGLES> -t 15 -k 21,33,55,77,99,127 -careful`. Gene features were identified using Prodigal (2.7) ([Hyatt et al., 2010](#)). Gene features were then annotated using Prokka 1.10 using the options: `-evalue 0.00001 -addgenes -kingdom Archaea -rfam -prefix<PREFIX> -cpus 4 -outdir <OUTDIR> <CONTIGS>` ([Seemann, 2014](#)). Ribosomal RNA genes were annotated using barrnap 0.5.2. To check for contamination, I mapped high-quality raw reads to the Greengenes database ([DeSantis et al., 2006](#)) and found that all reads mapping to 16S rRNA gene fell within the Euryarchaeota. The dynamic genome assessment tool CheckM was used to assess contamination on the basis of the conserved single copy protein domains ([Parks et al., 2015](#)). Duplicate single copy genes identified by CheckM were also checked manually.

Metagenome Binning

Illumina (HiSeq) shotgun genomic reads were screened against Illumina artifacts (adapters, DNA spike-ins) with a sliding window with a kmer size of 28 and a step size of 1. Reads with three or more ambiguous base calls or with average quality score of less than Q20 and a length <50 bps were removed. Screened reads were trimmed from both ends using a minimum quality cutoff of 5 using Sickle3. Trimmed, screened, paired-end Illumina reads were assembled using IDBA-UD ([Peng et al., 2012](#)) with the following parameters: `--pre_correction --mink 55 --maxk 95 --step 10 --seed_kmer 55`. To maximize assembly reads from different sites were co-assembled, as detailed in [Baker et al. \(2015, 2016\)](#) and [Lazar et al. \(2015\)](#). Initial binning of the assembled fragments was done using tetra-nucleotide frequencies signatures and emergent self-organizing mapping (ESOM) as detailed in [Dick et al. \(2009\)](#) and binning was enhanced by incorporating coverage signatures for the assembled contigs ([Sharon et al., 2013](#)). Binned contigs associated with the WOR_SM1_SCG were first identified via BLASTN alignments ([Altschul et al., 1990](#)). The completeness of the genomes within bins was then estimated using CheckM ([Parks et al., 2015](#)). Coverage was determined by recruiting reads to

scaffolds by BLASTN (bitscore > 75). Binning was also manually curated based on GC content, top blast hits, and mate-pairings.

Phylogenetics

Phylogenetic inference followed the methods of [Lloyd et al. \(2013\)](#). The position based gene homology search tool hmmer v3.1b2, Hidden Markov models based on Clusters of Orthologous Genes (COGs), and in-house scripts were combined to locate the 43 conserved genes within 114 archaeal genomes from every archaeal genus having genomic information available on IMG and assessed to be greater than 50% complete by CheckM ([Eddy, 2011](#); [Markowitz et al., 2014](#); [Parks et al., 2015](#); [Huerta-Cepas et al., 2016](#)). The alignment and tree building software ARB was used for trimming and to construct maximum likelihood trees, while bootstrap support for 16S rRNA gene trees were assessed using the RaxML rapid bootstrapping method. Additionally, Phylobayes was used to inference concatenated single copy gene trees using eight chains of “pb -d <trimmed_alignment.phy> -cat -gtr -dgam 4 chain<X>”. Chains were ended when the maxdiff was observed to be less than 0.3 using bpcomp -x 500 2 chain1...chain8. ArginyL-tRNA synthetase and nine conserved ribosomal genes (RplS7/L1/L3/L4/L2/L14/L16/S4E/L15) were used in this analysis. A subset of three of these genes present in candidate division pMC2A384 archaeon sp. SCGC AAA252-I15 was used to add the SAG to the tree via maximum parsimony. Sequences were aligned with mafft and trimmed using ARB as detailed previously ([Westram et al., 2011](#); [Lloyd et al., 2013](#)). A larger set of 26 conserved marker genes were also considered separately for genomes that contained them; alignments and trees were constructed using default settings in ETE2 ([Huerta-Cepas et al., 2010](#)). *In silico* primer alignments and annotation of functional gene alignments were done using the CLC Genomic Workbench 6.5.25.

Comparative Genomic Analysis

After gene calling and annotation a variety of bioinformatic toolsets were used to compare the seven genomes in the study (Table 2.1). Comparative genomics analysis was performed using methods detailed at using Anvi'o platform to visualize the output of ITEP 1.1 ([Benedict et al., 2014](#); [Eren et al., 2015](#)). Potential homologs across the genomes were discovered using the default BLAST alignment parameters within ITEP. Protein clusters were assigned by the Markov Cluster Algorithm (MCL) using the maxbit scoring method with the

inflation value set to 2.0 the score cutoff set to 0.4. Homologous gene sets for different groups of Altiarchaeales were determined based on similarity of genes within each subset of genomes, not the presence of a gene with a particular annotation. So, some genes with similar annotations can be present in multiple core sets, although they are different enough from each other that they are not considered homologous.

Hamus Protein Homolog Comparison

One protein (EMBL accession no. A0A098E857) found to be expressed in *Candidatus* Altiarchaeum hamiconexum ([Probst et al., 2014](#); [Perras et al., 2015](#)) was used as a query in a BLAST alignment to all predicted proteins from the six genomes listed in Table 2.1 as well as candidate division pMC2A384 archaeon sp. SCGC AAA252-I15 (pMC2A384) using default settings. Each alignment with *e*-value lower than 1×10^{-60} was considered as a possible homolog, because I found this value to adequately separate paralogs with similar domains to the hami proteins within the MSI genome. Similarly, BLAST alignments to the putative S_Layer_N domain region (aa position 5–81) in A0A098E857 and hidden Markov model based alignments to Euryarchaeotal S Layer domain proteins (ENOG410KVWW) were also conducted ([Altschul et al., 1990](#); [Huerta-Cepas et al., 2016](#)). Potentially homologous alignments were inspected manually in order to check for erroneous alignments to repeat regions and apparent pseudogenes. Then each putative protein was aligned against the UniProtKB database in order to visualize the alignment to conserved domain architectures.

Data Archiving

Short reads and assembly from WOR_SM1_SCG were submitted to NCBI BioProject PRJNA321288 (Accession Numbers: SRR3575064 and MCBE000000000). WOR_SM1_79 (Accession Number: MCBD000000000) and WOR_SM1_86-2 (Accession Number: MCBC000000000) were added to the NCBI BioProject PRJNA270657. IMC4_SM1 genome assembly was added to the NCBI BioProject PRJEB6121 (Accession Number: MCBF000000000). The MSI_SM1 genome assembly was previously made publicly available (Accession Number: CCXY000000000.1). CG_SM1 was previously available in the NCBI short read archive Acc. No. SRR1534154. Geochemical data are housed at www.bco-dmo.org.

Results

Geochemical Setting for Altiarchaeales

Sulfate diffusing across the sediment-water interface (6 mM) was mostly depleted by 61.5 cm into the sediments (Figure 2.2). Below this point, methane concentrations increased with depth in a concave up fashion suggesting anaerobic methane oxidation (Martens and Berner, 1977). Methane concentrations reached saturation (~1.2 mM) near the bottom of the core (70.5 cm; Figure 2.2). A sulfide smell was detected below about 9 cm, consistent with previous cores from this site. The good vertical alignment of the geochemistry in the two cores suggests that this geochemical setting was spatially stable, a result that has been noted previously ([Lloyd et al., 2011](#)).

In the WOR metagenomes, 16S rRNA genes for Altiarchaeales were only found at or below the sediment layer where downwardly diffusing SO_4^{2-} meets upwardly diffusing CH_4 (SMTZ). 16S rRNA gene sequences that are 100% identical to >900 bp of WOR_SM1_SCG were present in the SMTZ (16-24 cm) and the methanogenic zone (52–54 cm), but absent in the sulfate-rich zone (8–12 cm). The presence of Altiarchaeales at or below the SMTZ agrees with previous findings, which have only recovered similar 16S rRNA gene sequences from the methanogenic zone ([Kubo et al., 2012](#)).

Genome Quality Assessment

WOR_SM1_SCG assembled to a total size of 2.55 Mbp with 300 contigs above 1000 bp, the largest of which is 62,700 bp (Table 2.1). Contigs from WOR_SM1_79 and WOR_SM1_86-2 show significant similarity to the single cell genome assembly ([Altschul et al., 1990](#)). WOR_SM1_86-2 was taken from the SMTZ and WOR_SM1_79 taken from the methane-rich zone. WOR_SM1_79 and WOR_SM1_86-2 contain 361 and 170 contigs, totaling 3.19 and 2.09 Mb with maximum contig sizes of 93,451 and 80,876 bp, respectively.

WOR_SM1_SCG appeared to contain only one genome with a single copy of the 16S, 23S, and 5S rRNA gene sequences with predicted contamination rate of ~5% according to CheckM. Only five putatively single copy conserved genes were duplicated, and two of them shared a contig. Another pair matched each other 100% in the overlapping region, suggesting a sequence assembly error, while a fourth pair appeared to be the result of a mis-annotation of PFAM marker PF01287. The final pair was in PF03950, tRNA synthetase Class I, which was

duplicated in 189 out of the 4,227 bacterial and archaeal finished genomes on IMG suggesting these marker genes are naturally duplicated in a small subset of known genomes. Removing 29 contigs containing duplicate gene markers from WOR_SM1_79 and no contigs from WOR_SM1_86_2 reduced the contamination rate of both MAGs below the suggested 5% cutoff ([Parks et al., 2015](#)). Anvi'o was used to create a phylogenetic tree from the remaining 43 single copy conserved genes and to check of any additional signs of contamination (Figure 2.3) ([Eren et al., 2015](#)). Some single copy genes appeared in multiple copies due to apparent fission events (discussed below), while other incidents of multiple copies likely indicate the presence of multiple strains within the assemblies. Multiple copies of some genes involved in the replication of DNA also point to strain-level contamination (Supplemental Data Table 2.1). The assembled SAG was slightly more complete than the MAGs, based on the presence of single copy conserved genes (Table 2.1) ([Lloyd et al., 2013](#)). The completeness of the three genomes compares favorably with the highest values obtained in similar studies ([Wrighton et al., 2012](#); [Lloyd et al., 2013](#); [Rinke et al., 2013](#); [Swan et al., 2014](#)).

Phylogeny

16S rRNA genes are present in six genomes and is monophyletic in the candidate order Altiarchaeales (Figure 2.4A), branching deeply within the Euryarchaeota. Altiarchaeales also branch with the novel phylum DPANN ([Rinke et al., 2013](#)), although the low bootstrap support and variable placement with different phylogenetic tests suggest long-branch attraction. The concatenated subset of the 10 conserved genes present in the six genomes under consideration indicate that these genomes are monophyletic and distinct from the majority of the Euryarchaeota according to Bayesian and maximum likelihood analyses (Figure 2.4B). Candidate division pMC2A384 archaeon sp. SCGC AAA252-I15, derived from Sakinaw Lake ([Rinke et al., 2013](#)), was not given a taxonomic classification in its initial publication, since it lacked a 16S rRNA gene. According to our analysis, three single-copy conserved genes from that study overlap with 10 single-copy conserved genes (Figure 2.4B) and 23S rRNA gene sequences (data not shown) from this study causing pMC2A384 to branch with Altiarchaeales. These concatenated gene phylogenies agree with 16S rRNA gene phylogenies implying that Altiarchaeales is a monophyletic clade branching deeply within the Euryarchaeota, and are likewise susceptible to long-branch attraction with the DPANN.

Further evidence for the monophyly of Altiarchaeales comes from a putative phenylalanine tRNA synthetase beta subunit (*pheT*) that occurs in five of the six genomes in Table 2.1 and in candidate division pMC2A384 archaeon sp. SCGC AAA252-I15. In each genome, the *pheT* is split just before the B5 domain, which is a putative DNA-binding domain. Alignments of the split *pheT* to representatives from Archaea reveal that these genes share nearly all well conserved residues (Figure 2.5). Each half of the each split *pheT* gene occurs in the middle of assembled contigs, so the splits do not appear to be artifacts from the fragmented genome or an assembly error. The loci of the splits are well supported by reads that map to these contigs, and the splits are present across the different assemblers used for these genomes (IDBA_UD, SPAdes, and Mira). However, I could not determine their relative placements in the genomes, since the two pieces occur on different contigs in each genome. The position near the B5 domain within *pheT*, where the split occurs, shows poor homology across all three domains of life. *Homo sapiens*, *Saccharomyces cerevisiae*, *Helicobacter pylori*, as well as, several of the aligned archaeal *pheT* genes possess multiple amino acid insertions in the same location (Figure 2.4) ([Safro et al., 2004](#)); however, no currently sequenced genome has a split in *pheT* like I observe in our Altiarchaeales genomes. Additionally, one other split gene is found in all six Altiarchaeales genomes which is annotated as 2-(3-amino-3-carboxypropyl)histidine synthase. This gene appears to be homologous to and shares a similar split site with another split gene found in a Nanoarchaeota SAG collected from Obsidian Pool (Figure 2.6) ([Podar et al., 2013](#)). Within the Altiarchaeales, the seven genomes show clear and consistent subdivisions. Our analysis of 16S rRNA gene sequences groups WOR_SM1_SCG, both WOR_MAGs, and IMC4_SM1 into a one clade designated Alti-2, and CG_SM1 (USA, UT) and MSI_SM1 into another labeled Alti-1 (Figure 2.4A). Alti-2 contains the majority (88%) of 16S rRNA gene sequences from environmental libraries, including all other sequences from marine sediment. Additionally, one of the sequences in this clade was previously obtained from the same station in WOR sediments (JN605146; [Kubo et al., 2012](#)). This clade, however, is not marine/estuarine specific because it also contains 16S rRNA gene sequences derived from cold springs, hot springs, a hydroelectric dam, coastal and hypersaline aquifer, deep sea brines, and lacustrine sediments. In contrast, the clade containing the previously described genomes from CG_SM1 and MSI_SM1 contains only sequences from cold and hot springs. Mismatches to commonly

used 16S rRNA gene primers are found in all Altiarchaeales 16S rRNA gene sequences in this study, so it is likely that environmental abundances of Altiarchaeales have been masked in previous studies (Figure 2.6) ([Takai and Horikoshi, 2000](#); [Teske et al., 2002](#); [Yu et al., 2005](#)). Sequences from Alti-1 have 1 mismatch to the ARC-8F ([Teske et al., 2002](#)), while sequences from Alti-2 have 0 mismatches. Both Alti-1 and Alti-2 have an insertion within the sequence complementary to ARC-1492R ([Teske et al., 2002](#)). pMC2A384 appears to cluster with the other Alti-2 genomes; however, only a subset of amino acid positions with no 16S rRNA gene was available for the phylogenetic analyses used to make this inference. Within the Alti-2 clade, the 16S rRNA gene sequences from the WOR group together and separate from most other sequences, including those of IMC4_SM1. WOR_SM1_SCG has high 16S rRNA gene sequence identity with the other sequence found previously in WOR sediments, JN605146 and WOR_SM1_79 (95.6 and 95.3%, respectively). WOR_SM1_86-2 shares only 83.1% 16S rRNA gene sequence identity to WOR_SM1_SCG with no closer sequences available from the NCBI database. WOR_SM1_SCG and WOR_SM1_79 rRNA gene sequences cluster with other sequences from marine sediments, and IMC4_SM1 clusters with other sequences from sulfidic springs (Figure 2.4A).

Comparative Genomics of the Altiarchaeales

Only 57 genes are homologous across all seven genomes in this study (Figure 2.7). However, 573 genes are shared between at least some members of the distantly related Alti-1 and Alti-2 genomes, and they referred to here after as the “core Altiarchaeales” genes. Although many of these are conserved universally, some of them may be more specific to Altiarchaeales. These include genes associated with oxidative stress, glycolysis, the TCA cycle, transcription/translation, transporters, ATP synthase, pyruvate synthase, acetyl-CoA synthase, F₄₂₀ dehydrogenase, ferredoxin and biosynthesis of amino acids, sugars, lipids, and cobalamins. Also included in the core Altiarchaeales genome are key genes in the Wood–Ljungdahl pathway for CO₂ fixation such as 5,10-methylenetetrahydrofolate reductase, Acetyl-CoA decarbonylase/synthase complex and carbonic anhydrase, and NADP-dependent methylenetetrahydromethanopterin dehydrogenase. All these Wood–Ljungdahl pathway genes are similar to genes found in archaea, except NADP-dependent methylenetetrahydromethanopterin dehydrogenase, which are more commonly found in some.

Table 1. Statistics on genomes highlighted in this study.

	Freshwater Altiarchaeales				Estuarine Altiarchaeales	
	MSI_SM1	CG_SM1	IM-C4_SM1	WOR_SM1_SCG	WOR_SM1_86-2	WOR_SM1_79
Status	Draft	Draft	Draft	Draft	Draft	Draft
Contigs	467	220	1	300	170	332
basepairs (Mb)	3.33	1.46	1.39	2.55	2.09	3.19
GC content (%)	32.09	32.73	48.48	37.12	40.83	39.76
N50	7996	8841	1390915	16971	15808	10683
Predicted proteins	3005	1388	1478	2514	2297	3275
tRNA	70	37	18	23	23	24
Completeness ^a	95	90	79	81	77	58
Reference	Probst et al, 2014	Probst et al, 2014	Probst et al, 2014	This Study	This Study	This Study

^aCalculated by CheckM, accounting for the number of nearly-universal conserved genes present.

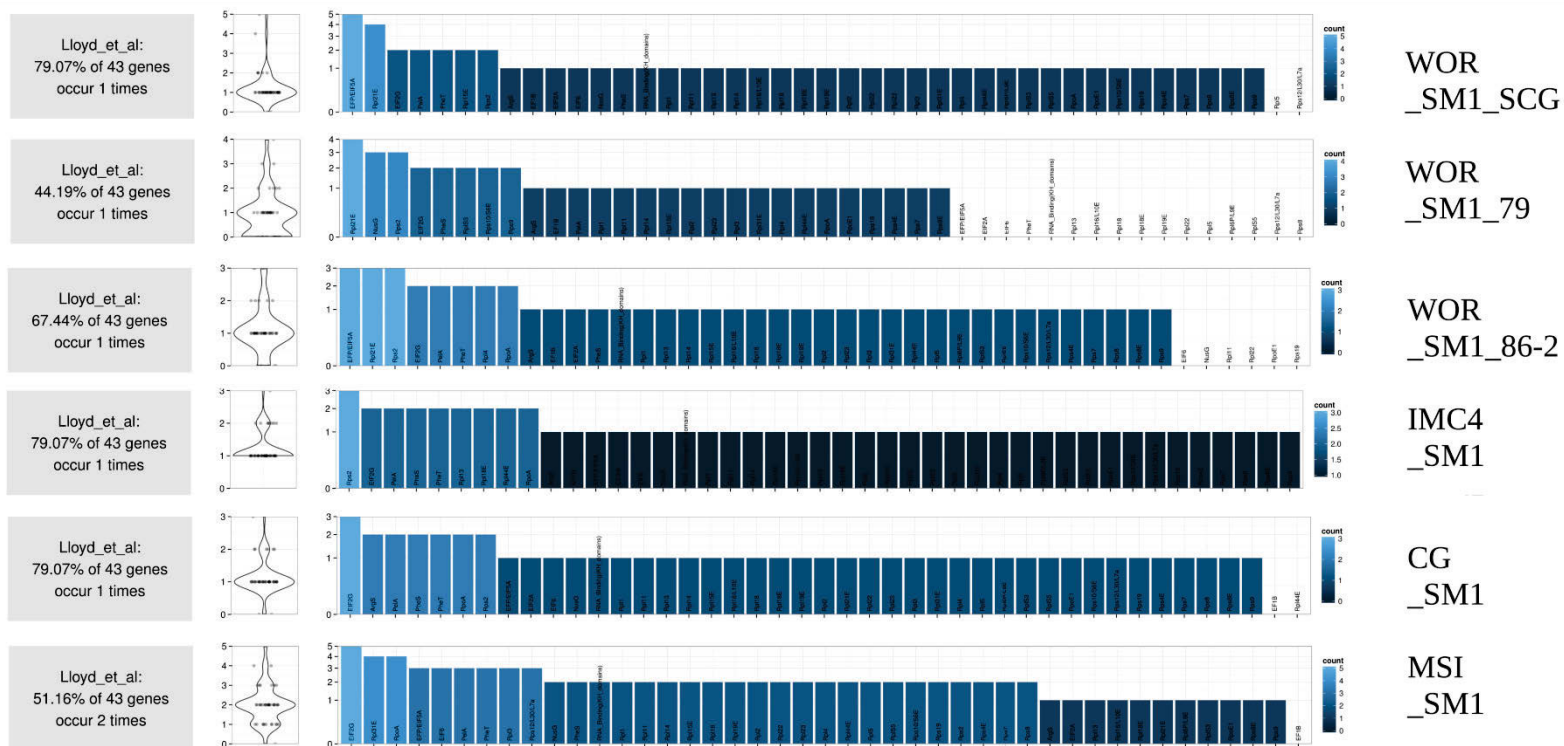


Figure 2.3: Single copy conserved genes in genomes. Histograms are labeled with gene names and the y-axis depict the number of positive hits to hmmer models in set of 43 single-copy genes which are conserved in most archaea ([Lloyd et al., 2013](#)). The 1-D plot summarize the data and the mode is used to estimate the number of genomes in the assembly.

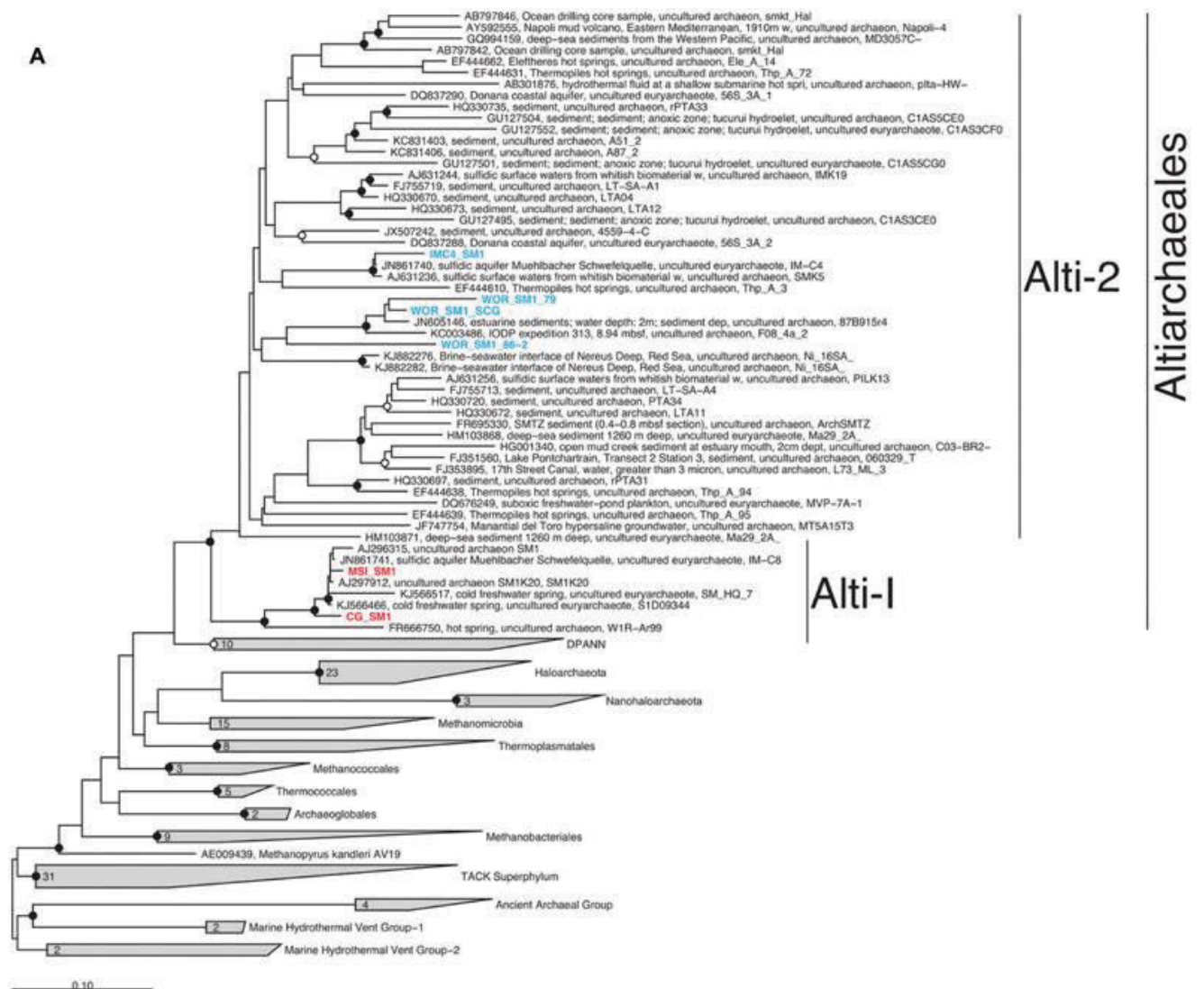


Figure 2.4A: Phylogenetic placement of Altiarchaeales based on (A) a maximum likelihood tree constructed from full length (>1300 bp) 16S rRNA gene sequences; smaller sequences (>900) within the Altiarchaeales group were added to the tree using maximum parsimony (ARB); filled and open circles indicate bootstrap support at greater than 90 and 70% respectively, ...

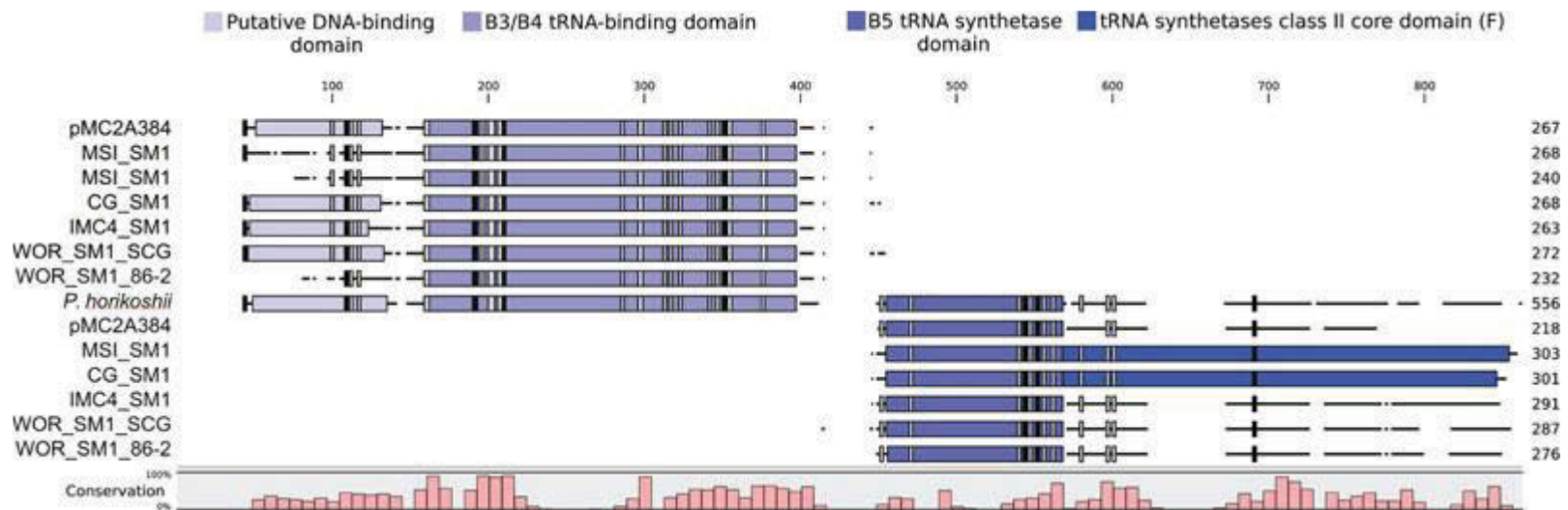


Figure 2.5: The amino acids of phenylalanine tRNA-synthetase beta subunit (*pheT*) in six proposed Altiarchaeales genomes were aligned using ClustalO to 104 *pheT* sequences from one representative of each archaeal genus with draft or finished genomes above 50% completeness, but only *pheT* from *Pyrococcus horikoshii* is shown. The protein alignment is visualized along the x-axis and labeled by genome name along the y-axis. Colored boxes placed over the aligned sequences represent protein domains associated with *pheT* (InterProScan5). Black and gray boxes denote the presence of conserved residues found in 100 or 90% of archaeal *pheT*, respectively. The histogram tracks the relative conservation of residues across all aligned archaeal *pheT*.

types of bacteria. The core Altiarchaeales genome also contains 69 hypothetical proteins. The core Alti-1 genome, represented by the spring-associated MSI_SM1 and CG_SM1, has 436 shared genes besides those in all the Altiarchaeales genomes. These included glycosyltransferase, lipoprotein *nlpI*, folate synthesis genes, four CRISPR-associated genes, a Type III restriction endonuclease, and two Type I restriction endonucleases. Two homologs to glycotransferases, *rfaB*, are found only in the Alti_1 set. Putative polysaccharide synthesis genes, which are thought to be important for biofilm production, are exclusive to the Alti-1 set ([Probst et al., 2014](#)). Colanic acid export protein, *wza*, is only found in MSI_SM1. 150 hypothetical proteins are also found in the core Alti-1 set (Supplemental Data Table 2.1). The core Alti-2 genome, deduced from five genomes, share 836 homologous genes including archaeal ATP synthase, cyclic pyranopterin monophosphate synthase, adenylate kinase, flagellar proteins, hydrogenases, and mechanosensitive channels. With the exception of the freshwater member of Alti-2, IMC4_SM1, the shared genes also include sodium specific transporters, calcium binding proteins, sulfite exporter, sulfoxide reductase, adenylylsulfate kinase, and Co-A disulfide reductase.

Comparison of Putative Hami Proteins

Homologs to the full-length hamus protein from MSI_SM1 are not found in the six other genomes in this study. The hamus protein in MSI_SM1 (EMBL accession no. A0A098E857) is 1808 aa in length and contains three PFAM domains (S_Layer_N, NosD, FKBP_C). A putative homolog, CG_1237, which is significantly truncated (372 aa vs. 1808), lacks an S_Layer_N domain but contains the FKBP_C domain. IMC4_884 is also significantly truncated (918 aa vs. 1808), but it is the only putative homolog that contains a putative S_Layer_N domain. Additionally, like the hamus protein found in MSI_SM1, it is predicted to be localized to the membrane containing an N terminal signal sequence and a predicted C terminal trans membrane domain ([Möller et al., 2001](#); [Perras et al., 2015](#)). A putative homolog, SCG_1643, which contained FKBP_C and Thioredoxin_4 domains, is more similar to another protein in MSI_SM1, a putative peptidylprolyl isomerase (EMBL A0A098E9R9), and is also much shorter at 613 aa. WOR_SM1_79 has another potential homolog, 79_2061, but this was very similar to SCG_1643, with alignments of the two only having 10 amino acid changes with no gaps. 86-2_679 has no

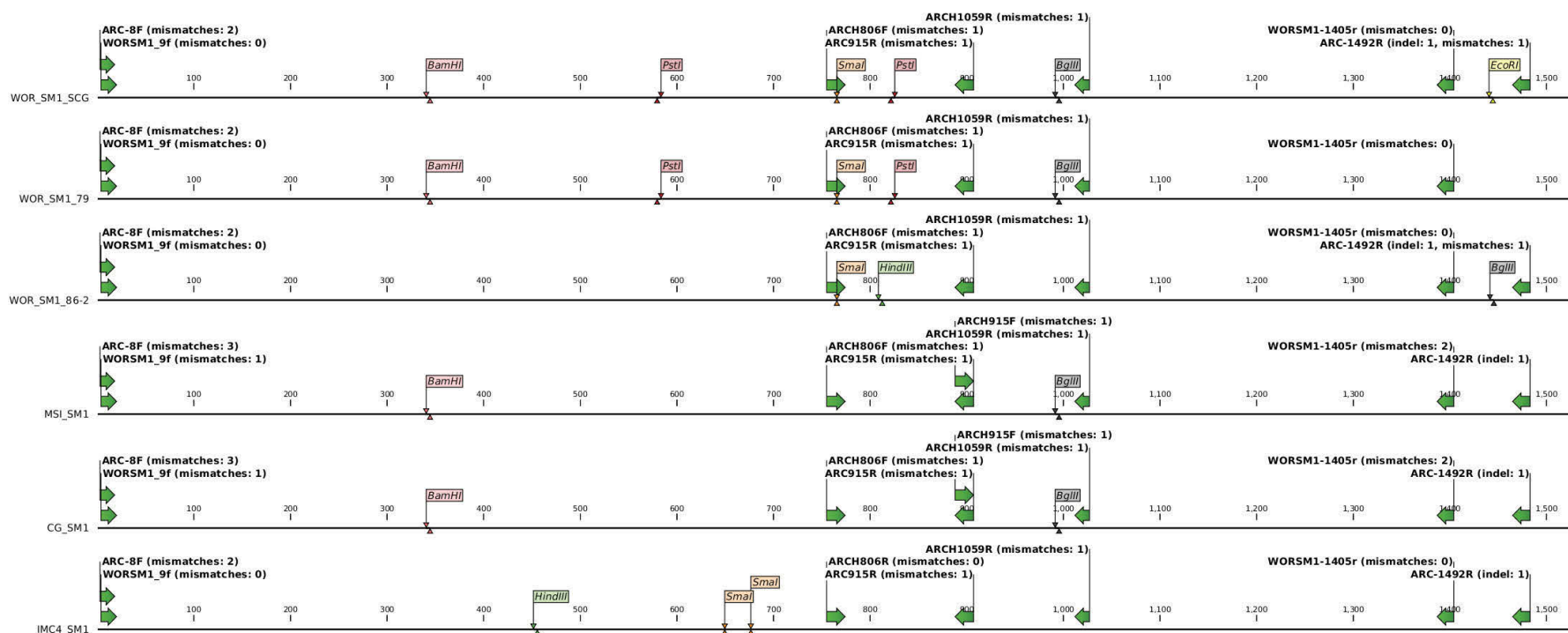


Figure 2.6: Mismatches in common full length archaeal primers. Primer sequence alignment and predicted restriction digestion sites are annotated on the full length 16S rRNA gene sequences. Green annotations were assigned by the primer alignment software (CLC Genomics Workbench 6.5.2). Ambiguous base calling prevented the alignment of ARC-1492R to WOR_SM1_79. WORSM1-9f (5'-CCAGTTGATCCTGCTGG-3') and WORSM1-1405r (5'-CCACTCGATTGGTTTGAC-3') were designed for this study.

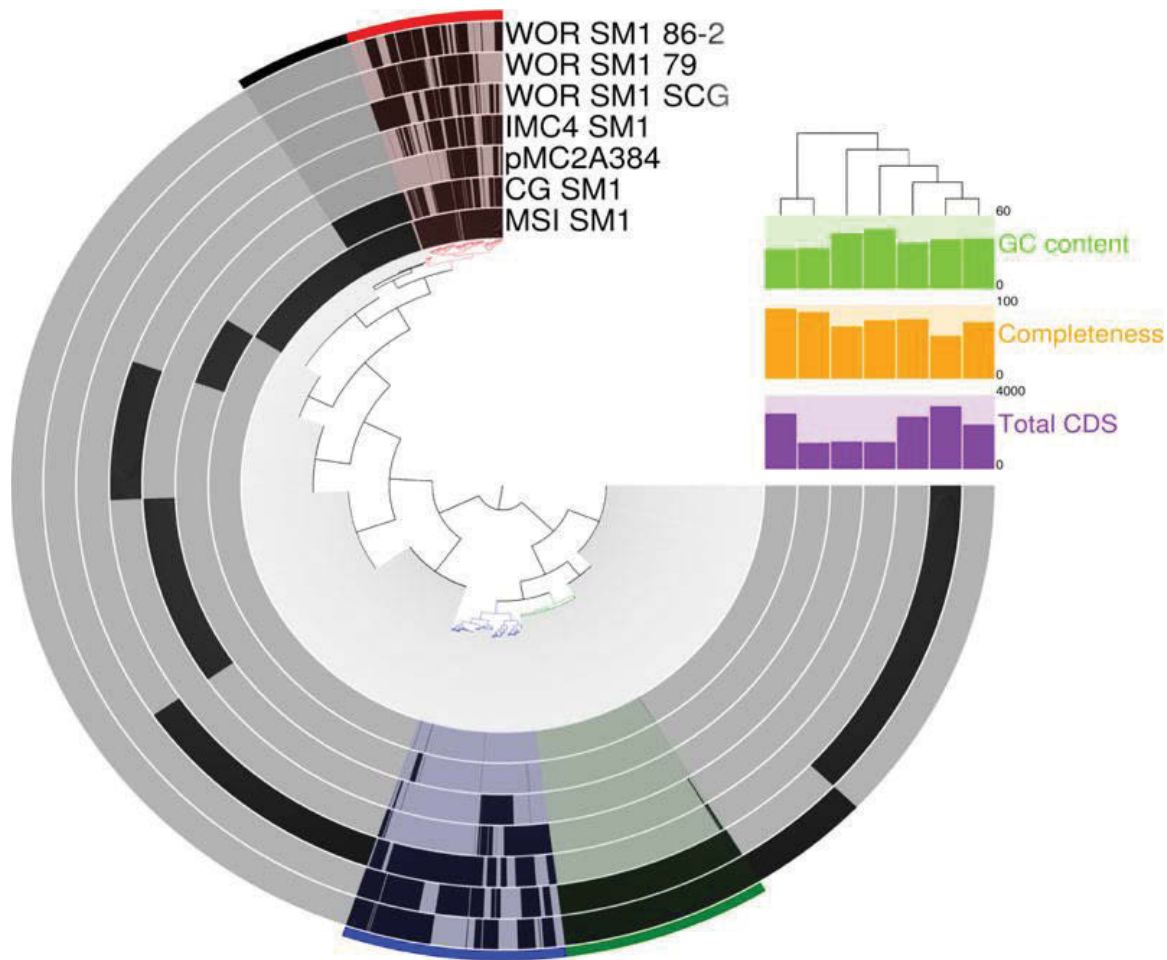


Figure 2.7: Comparative analysis of protein clusters within all known Altiarchaeales genomes. The inner tree was constructed from matrix of protein abundances across the distantly related genomes in this study. Black bars at the tips of inner tree represent the presence or absences of a protein within a genome. Colored bins along the edge highlight co-occurrence patterns across the genomes: core Altiarchaeales proteins (red), core Alti-1 (black), core Alti-2 (blue), and homologous proteins between WOR genomes from metagenomes (green). G + C (bright green), predicted completeness (orange), and total predicted coding sequences (purple) are displaying to the right. Branch lengths in the tree above the histograms correspond to dissimilarities between predicted coding sequences in the genome.

similar PFAM domains to A0A098E857, yet it also has no other significant matches within the UniProtKB database.

Discussion

Phylogeny of the Altiarchaeales

[Rudolph et al. \(2001\)](#) first reported a conspicuous cold-loving biofilm-forming archaea. They named SM1 after the Bavarian spring Sippenauer Moor. Subsequent studies have revealed a broad diversity of 16S rRNA gene sequences distantly related to those organisms ([Rudolph et al., 2004](#); [Probst and Moissl-Eichinger, 2015](#)), most of which were found in environments with no conspicuous mats (Figure 2.1). By including genomes from estuarine sediments in my phylogenetic analyses, I was able to identify two major groups within the Altiarchaeales, Alti-1 and Alti-2. Alti-1 is composed primarily of sequences isolated from spring environments, while Alti-2 sequences are from a variety of primarily anoxic environments. I show that only sequences from environments with actively moving water fall into the relatively low-diversity sister clade Alti-1. This clade is distinct from the more widespread, diverse group of Alti-2 as previous 16S rRNA gene surveys have suggested ([Probst et al., 2014](#); [Probst and Moissl-Eichinger, 2015](#)). These data should be interpreted with caution as mismatches in commonly used 16S rRNA gene primers have likely left the diversity of either clade in a given environment under-sampled. Although IMC4_SM1, a member of the Alti-2, was extracted from the same sample as MSI_SM1, it was shown by qPCR and fluorescence *in situ* hybridization (FISH) to be in low abundance within the biofilm ([Probst et al., 2014](#)). Given the large distance between Alti-1 and Alti-2 (>20% dissimilarity by 16S rRNA genes), I sought to determine whether these two clades are truly monophyletic. Our 16S rRNA gene tree shows good statistical support for the monophyly of the Altiarchaeales sister clades, in agreement with previous work ([Probst et al., 2014](#)). However, single copy gene trees from whole genomic constructions have been hindered by lack of representation of genomes from the more widespread Alti-2. The inclusion of our three WOR genomes and one from Sakinaw Lake in this study supplement the previously known IMC4_SM1 bringing the representation from Alti-2 to five genomes. These additional genomes have good posterior probability support for the conclusion that the monophyletic lineage, Altiarchaeales, is composed of two distinct clades, Alti-1 and Alti-2.

Further support for the monophyly of these clades comes from shared traits encoded by the genomes in this study. The acquisition or loss of genes involved in genome replication have been shown to be helpful in describing the shared evolutionary histories of archaea phyla ([Raymann et al., 2014](#)). All Alti-1 and Alti-2 genomes have the same set of replication genes (Table 2.2) ([Probst et al., 2014](#)). In addition to shared replication machinery, my analysis points to synapomorphic mutations in the phenylalanine tRNA beta subunit *pheT* of six of seven genomes and in the 2-(3-amino-3-carboxypropyl)histidine synthase of all seven genomes. Although, gene fissions are rare events that can be used to support phylogenies, shared split genes in Nanoarchaeota have been used to support a relationship between terrestrial and marine hyperthermophilic biotypes despite the genome plasticity that is commonly associated with parasitic lifestyles ([Waters et al., 2003](#); [Podar et al., 2013](#)). *pheT* and 2-(3-amino-3-carboxypropyl)histidine synthase are present in a single copy in nearly all other archaea. The fact that they are split in Altiarchaeales suggests that apparent duplications of these genes reported in this and previous work may actually be fission events (Figure 2.5) ([Probst et al., 2016](#)). Based on this evidence and the placement of candidate division pMC2A384 archaeon sp. SCGC AAA252-I15 within Figure 2.4B, I propose that pMC2A384 belongs within the larger Altiarchaeales classification. There is precedent for split tRNA synthase genes in *Aquifex aeolicus*, whose leucine-tRNA synthetase gene is split into two pieces at the tRNA-binding domain site ([Ma et al., 2006](#)) and Nanoarchaeota which have split Glu-tRNA_{Gln} amidotransferase and Alanyl-tRNA synthase ([Waters et al., 2003](#); [Podar et al., 2013](#)). While a similar split 2-(3-amino-3-carboxypropyl)histidine synthase was observed previously in a single Nanoarchaeota SAG, the lack of this split in other Nanoarchaeota genomes and the lack of phylogenetic evidence linking Nanoarchaeota to Altiarchaeales lead us to concluded that these represent two separate evolutionary events. Protein fission events have been well documented in the literature as a mechanism of new gene formation in prokaryotes ([Snel et al., 2000](#); [Suhre and Claverie, 2004](#); [Kummerfeld and Teichmann, 2005](#); [Pasek et al., 2006](#)). The function of this split cannot be determined from the primary sequence, but it is likely to be related to an evolutionary event occurring in an ancestor of these organisms that produced a novel variant of the universally conserved phenylalanine tRNA-synthetase beta subunit. Additionally, the split occurs between two protein domains and nearly all conserved residues are present (Figure 2.5) suggesting that

the split *pheT* gene is functional and recombines post-translationally. This apparent gene-splitting event appears to be a unique character that helps to define this candidate order. Our results partially agree with placement of Altiarchaeales within the Euryarchaeota ([Probst et al., 2014](#)). It should be noted that our inferences include members of the candidate phylum Woesearchaeota ([Castelle et al., 2015](#)) that were not available during earlier analyses ([Probst et al., 2014](#)). Our results disagree with previous results showing robust placement of Altiarchaeales (labeled Altiarchaeota in that paper) within the DPANN superphylum in a clade that includes another archaeon with a double-membrane, the ARMAN group ([Seitz et al., 2016](#)). It is important to note that, while DPANN and Altiarchaeales appear as sister groups in Figure 2.4, I found no statistical support for this branching pattern, suggesting this is a long branch attraction artifact. Finally, each of these results differ from recent work which suggests that Altiarchaeales could be a separate phylum-level lineage distinct from either the DPANN and Euryarchaeota ([Hug et al., 2016](#)). Six of the 10 marker genes used in concatenated gene trees in Figure 2.4B overlap with the 16 marker genes used in two of these previous works ([Hug et al., 2016](#); [Seitz et al., 2016](#)). Such wide disagreement between tree topologies warrants further study and suggests Altiarchaeales is quite distantly related to all other taxa studied to date. Our phylogenetic and trait-based analysis suggest that the genomes in this study belong to two distinct, yet monophyletic clades which may belong within a higher taxonomic classification than previously suggested; although, for now, there is no conclusive evidence that they are not members of Euryarchaeota ([Probst et al., 2014](#)).

When comparing the tree topologies between 16S rRNA gene and protein trees within Alti-2, I find some inconsistencies. 16S rRNA gene sequences suggest WOR_SM1_SCG is most closely related to WOR_SM1_79, while protein sequences suggest IMC4_SM1 shares a more recent ancestor with WOR_SM1_SCG. In contrast to the ambiguous position within the Alti-2 for WOR_SM1_86-2 in 16S rRNA gene analyses, protein sequence trees suggest WOR_SM1_86-2 is very closely related to WOR_SM1_79. Data from 16S rRNA genes obtained from metagenomes should be interpreted with caution, as it is often difficult to bin and assemble 16S rRNA gene sequences because of the presence of highly conserved regions and divergent genomic signatures.

Genome Comparisons within the Altiarchaeales

Variation in Genome Size

Across the Altiarchaeales genomes there was considerable variation in GC content and projected genome size. Previous analysis of synonymous vs. non-synonymous mutations within housekeeping genes suggests the MSI_SM1 draft genome represents multiple strains of *Candidatus* Altiarchaeum hamiconexum ([Probst et al., 2014](#)). Analysis of housekeeping genes within the other draft genomes presented here also reveals the potential for contamination from multiple strains to a lesser extent. Draft genomes from the WOR are larger than those from terrestrial spring; however, the lack of complete genomes and the considerable evolutionary distance between the genomes impairs our ability to draw conclusions about environmental pressures which may influence genome size and GC content.

Genomic Evidence for Hami

Genomes from the Alti-1 and Alti-2 were compared to establish functional traits that help to define the group as a whole as well as to identify potential functions useful within their particular environments. Since the grappling-hook-like hami are a defining physical characteristic of *Candidatus* Altiarchaeum hamiconexum, I investigated the other genomes to determine whether this feature is characteristic of all Altiarchaeales. The hamus structure likely diverged from ancient S-layer proteins following the acquisition of a double membrane structure([Perras et al., 2015](#)). I find that there is evidence for the hamus protein in the IMC4_SM1 genome of the Alti-2 clade; however, future work should seek to determine if these proteins are expressed *in situ*. No such homologs containing S_Layer_N domains are observed in CG_SM1; although, new evidence suggests CG_SM1 nonetheless have hook-like appendages (A. Klingl, personal communication). None of the genomes from the WOR possess hami protein homologs with S_Layer_N domains. This points to the adaptive utility of hami in the actively flowing stream environment. Given the dramatic differences in length and domain architecture in hami protein homologs, future studies should highlight the diversity of form among hami ultrastructures in the Altiarchaeales.

Biofilm Production

The Alti-1 core genome contains other genes possibly associated with biofilm formation that are absent in the Alti-2 core genome. The two Alti-1 specific glycosylases have similar domain architecture and homology to *rfaB*, the gene for UDP-

Table 2.2: Shared replication machinery across Altiarchaeales genomes. Each indicates an additional copy of the gene found in the genome assembly, while # indicates an additional homolog that appears truncated.

Genome	DNA Ligase	Cdc6/Orc	PoID-S	FEN-1	DNA Gyrase A	DNA Gyrase B	MCM	PCNA	Pol B	PrIS	PrIL	RFC-S	RnaseH II	TopoVI A	TopoVI B	RPAI
MSI_SM1	X	X**	X*#	X	X	X	X**	X	X**	X	X	X*	X*	X	X	X
CG_SM1	X	X	X	X	X	?	X	X	X	X	X	X	X	X	X	X
IMC4_SM1	X	X	X	X	?	X	X	X	X	X	X	?	X	X	X	X
WOR_SM1_SCG	X	X	X	?	X	X*#	X	X*	X*	X	X*	X	?	X	X	?
WOR_SM1_79	X	X	X**	X##	X**	?	X***	X	X*	X*	X	X	?	X**	X	X
WOR_SM1_86-2	X	X	X*	?	X**	X*	X	?	X*	X*	X	X	X*	X	X	?
pMC2A384	X	X	X	X#	?	?	X	?	?	X	X	X	X	X	X	?

galactose:(glucosyl)lipopolysaccharide-1,6-D-galactosyltransferase. This gene has been linked to lipopolysaccharide (LPS) synthesis in *E. coli*, and its mutants appear impaired in UDP-galactose and UDP-glucose catabolism ([Qian et al., 2014](#)). UDP-galactose catabolism is also required for *Bacillus subtilis* to survive in a biofilm, but is not required for planktonic growth ([Chai et al., 2012](#)). When lipoprotein NlpI, found only in Alti-1 genomes, is disabled in *E. coli*, the cells have a 35-fold decrease in their ability to adhere to intestinal lining ([Barnich et al., 2004](#)). This, in combination with the common participation of lipoproteins in biofilm formation ([Raaijmakers et al., 2006](#)), suggests that these genes are important for biofilm formation in Alti-1. Homologs of colanic acid export proteins *wza* as well as putative polysaccharide synthesis genes, which are thought to be important for biofilm production ([Probst et al., 2014](#)), are only found in MSI_SM1. Mutants for the *wza* enzyme in a *Klebsiella pneumonia* model system have been found to be deficient in producing biofilms ([Balestrino et al., 2008](#); [Wu et al., 2011](#)). While Altiarchaeales species are the dominant organism in Crystal Geyser spring water, biofilms similar to those seen in the German sulfidic springs have never been reported there ([Probst et al., 2014, 2016](#)). Biofilm formation may be an important trait for some Alti-1 members, while in other environments Alti-1 may use other mechanisms to meet the challenges of living within actively flowing streams.

Potential Adaptations and Metabolic Similarities

The core genome of the Alti-2 clade suggests adaptations to a free-living state in sulfidic environments. A few of the genes that differentiate them from Alti-1 are important to energy maintenance, such as ATP synthase, hydrogenases, and adenylate kinase which modulates intracellular ADP:ATP ratios. The presence of flagellar genes suggests that motility may be important in the free-living state. Quite a few Alti-2 genes suggest adaptations to an estuarine environment. This evidence is bolstered by the fact that these adaptations are absent in IMC4_SM1, the one Alti-2 genome from a freshwater environment. These include osmoregulation mechanisms such as sodium transporters and mechanosensitive channels. In addition, Alti-2 seems capable of sulfur metabolism, since it contains many sulfur-related genes. The presence of these genes could be for sulfur assimilation, but many of them involved in energy-conserving redox reactions. However, no dissimilatory sulfite reductase genes are found in this study. When one of the sulfur-related genes, CoA-disulfide reductase, is knocked out in *Pyrococcus furiosus* growing in the presence of sulfur compounds, the organism loses its ability

to survive transient exposure to oxygen ([Herwald et al., 2013](#)). Therefore, Altiarchaeales in the WOR may shuttle electrons between sulfur compounds for oxidative stress defense.

Despite all the phylogenetic and functional differences between Alti-1 and Alti-2, there are key similarities, namely the presence of many elements of glycolysis, the tricarboxylic acid cycle, cobalamin production/recycling, and all the genes in the Wood–Ljungdahl pathway for autotrophy, as well as, a lack of convincing evidence for catabolism of a broad range of heterotrophic substrates. Alti-1 has previously been suggested to have acquired its NADP-dependent methylenetetrahydromethanopterin dehydrogenase gene from a bacterium through horizontal gene transfer ([Probst et al., 2014](#)). The presence of this gene, also with a potentially bacterial origin, in our Alti-2 genomes supports the possibility that the modified Wood–Ljungdahl pathway seen in *Candidatus* Altiarchaeum hamiconexum is conserved in the Altiarchaeales.

Conclusion

The three novel genome reconstructions presented here belong to the monophyletic archaeal lineage Altiarchaeales, which can be divided into two clades, Alti-1 and Alti-2. Moreover, these genome reconstructions belong to the more diverse and widespread Alti-2 clade. Our analyses support previous claims that Altiarchaeales is a deeply branching order within the Euryarchaeota, but further research on deeply branching archaeal lineages is necessary to resolve divergences in tree topologies with the recent expansion of the DPANN superphylum. The core genes shared by all known genomes have a characteristic split *pheT* gene and genes for a version of the Wood–Ljungdahl pathway. However, Alti-1 and Alti-2 had genomic features that distinguished them from each other and suggest adaptations to different environments. Alti-1 have adaptations to form hooks in high flow environments and may form biofilms, while the members of Alti-2 inhabiting sulfidic, saline environments have adaptations that make them specialized for a free-living state.

Funding

This work was funded by NSF IGERT: SCALE-IT (0801540) (JB), NSF OCE-1431598 (KL, JB), and is NSF Center for Dark Energy Biosphere Investigations (OCE-0939564) contribution # 332 (KL, JB). Work at UCB was performed under the DFG grant PR1603/1-1 given to AP. MP was funded by the U.S. Department of Energy, Office of Biological and Environmental Research

(DE-SC0006654). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

Acknowledgments

I would like to thank Michael Piehler at the University of North Carolina Institute of Marine Sciences and Frank Loeffler at the University of Tennessee for their generosity in sharing lab space and equipment, Christine Moissl-Eichinger for providing data associated with the IMC4_SM1 genome, Captain John Baldrige for the use of his boat and time to access the sampling site, Steven A. Higgins whose scripts were adapted to aid in phylogenetic analyses, and Steve Allman for assistance with cell sorting.

List of References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Baker, B. J., Lazar, C. S., Teske, A. P., and Dick, G. J. (2015). Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* 3, 14. doi: 10.1186/s40168-015-0077-6
- Baker, B. J., Saw, J. H., Lind, A. E., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., et al. (2016). Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nat. Microbiol.* 1, 16002. doi: 10.1038/nmicrobiol.2016.2
- Balestrino, D., Ghigo, J.-M., Charbonnel, N., Haagsen, J. A. J., and Forestier, C. (2008). The characterization of functions involved in the establishment and maturation of *Klebsiella pneumoniae* in vitro biofilm reveals dual roles for surface exopolysaccharides. *Environ. Microbiol.* 10, 685–701. doi: 10.1111/j.1462-2920.2007.01491.x
- Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., et al. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* 25, 690–701. doi: 10.1016/j.cub.2015.01.014
- Chai, Y., Beauregard, P. B., Vlamakis, H., Losick, R., and Kolter, R. (2012). Galactose metabolism plays a crucial role in biofilm formation by *Bacillus subtilis*. *MBio* 3:e00184-12. doi: 10.1128/mBio.00184-12
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05
- Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10:R85. doi: 10.1186/gb-2009-10-8-r85
- Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

- Eren, A. M., Esen, ÖC., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:e1319. doi: 10.7717/peerj.1319
- Herwald, S., Liu, A. Y., Zhu, B. E., Sea, K. W., Lopez, K. M., Sazinsky, M. H., et al. (2013). Structure and substrate specificity of the pyrococcal coenzyme A disulfide reductases/polysulfide reductases (CoADR/Psr): implications for S₀-based respiration and a sulfur-dependent antioxidant system in *Pyrococcus*. *Biochemistry* 52, 2764–2773. doi: 10.1021/bi3014399
- Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24. doi: 10.1186/1471-2105-11-24
- Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44, D286–D293. doi: 10.1093/nar/gkv1248
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi: 10.1038/nmicrobiol.2016.48
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119
- Kelley, C. A., Martens, C. S., and Chanton, J. P. (1990). Variations in sedimentary carbon remineralization rates in the White Oak River estuary, North Carolina. *Limnol. Oceanogr.* 35, 372–383. doi: 10.4319/lo.1990.35.2.0372
- Kubo, K., Lloyd, K. G., Biddle, J. F., Amann, R., Teske, A., and Knittel, K. (2012). Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J.* 6, 1949–1965. doi: 10.1038/ismej.2012.37
- Kummerfeld, S. K., and Teichmann, S. A. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* 21, 25–30. doi: 10.1016/j.tig.2004.11.007
- Lazar, C. S., Baker, B. J., Seitz, K., Hyde, A. S., Dick, G. J., Hinrichs, K.-U., et al. (2015). Genomic evidence for distinct carbon substrate preferences and ecological niches of

- Bathyarchaeota in estuarine sediments. *Environ. Microbiol.* 18, 1200–1211. doi: 10.1111/1462-2920.13142
- Lloyd, K. G., Alperin, M. J., and Teske, A. (2011). Environmental evidence for net methane production and oxidation in putative ANaerobic Methanotrophic (ANME) archaea. *Environ. Microbiol.* 13, 2548–2564. doi: 10.1111/j.1462-2920.2011.02526.x
- Lloyd, K. G., Schreiber, L., Petersen, D. G., Kjeldsen, K. U., Lever, M. A., Steen, A. D., et al. (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* 496, 215–218. doi: 10.1038/nature12033
- Ma, J.-J., Zhao, M.-W., and Wang, E.-D. (2006). Split leucine-specific domain of leucyl-tRNA synthetase from the hyperthermophilic bacterium *Aquifex aeolicus*. *Biochemistry* 45, 14809–14816. doi: 10.1021/bi061026r
- Markowitz, V. M., Chen, I.-M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* 42, D568–D573. doi: 10.1093/nar/gkt919
- Martens, C. S., and Berner, R. A. (1977). Interstitial water chemistry of anoxic long island sound sediments. 1. Dissolved gases. *Limnol. Oceanogr.* 22, 10–25. doi: 10.4319/lo.1977.22.1.0010
- Martens, C. S., and Goldhaber, M. B. (1978). Early diagenesis in transitional sedimentary environments of the White Oak River Estuary, North Carolina. *Limnol. Oceanogr.* 23, 428–441. doi: 10.4319/lo.1978.23.3.0428
- Michaelis, W., Seifert, R., Nauhaus, K., Treude, T., Thiel, V., Blumenberg, M., et al. (2002). Microbial reefs in the Black Sea fueled by anaerobic oxidation of methane. *Science* 297, 1013–1015. doi: 10.1126/science.1072502
- Moissl, C., Rachel, R., Briegel, A., Engelhardt, H., and Huber, R. (2005). The unique structure of archaeal ‘hami’, highly complex cell appendages with nano-grappling hooks. *Mol. Microbiol.* 56, 361–370. doi: 10.1111/j.1365-2958.2005.04294.x
- Moissl, C., Rudolph, C., and Huber, R. (2002). Natural communities of novel archaea and bacteria with a string-of-pearls-like morphology: molecular analysis of the bacterial partners. *Appl. Environ. Microbiol.* 68, 933–937. doi: 10.1128/AEM.68.2.933-937.2002
- Moissl, C., Rudolph, C., Rachel, R., Koch, M., and Huber, R. (2003). In situ growth of the novel SM1 euryarchaeon from a string-of-pearls-like microbial community in its cold biotope, its

- physical separation and insights into its structure and physiology. *Arch. Microbiol.* 180, 211–217. doi: 10.1007/s00203-003-0580-1
- Möller, S., Croning, M. D. R., and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653. doi: 10.1093/bioinformatics/17.7.646
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114
- Pasek, S., Risler, J.-L., and Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22, 1418–1423. doi: 10.1093/bioinformatics/btl135
- Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174
- Perras, A. K., Daum, B., Ziegler, C., Takahashi, L. K., Ahmed, M., Wanner, G., et al. (2015). S-layers at second glance? Altiarchaeal grappling hooks (hami) resemble archaeal S-layer proteins in structure and sequence. *Front. Microbiol.* 6:543. doi: 10.3389/fmicb.2015.00543
- Perras, A. K., Wanner, G., Klingl, A., Mora, M., Auerbach, A. K., Heinz, V., et al. (2014). Grappling archaea: ultrastructural analyses of an uncultivated, cold-loving archaeon, and its biofilm. *Front. Microbiol.* 5:397. doi: 10.3389/fmicb.2014.00397
- Podar, M., Makarova, K. S., Graham, D. E., Wolf, Y. I., Koonin, E. V., and Reysenbach, A.-L. (2013). Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol. Direct* 8, 9. doi: 10.1186/1745-6150-8-9
- Probst, A., and Moissl-Eichinger, C. (2015). “Altiarchaeales”: uncultivated archaea from the subsurface. *Life* 5, 1381–1395. doi: 10.3390/life5021381
- Probst, A. J., Castelle, C. J., Singh, A., Brown, C. T., Anantharaman, K., Sharon, I., et al. (2016). Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations. *Environ. Microbiol.* doi: 10.1111/1462-2920.13362 [Epub ahead of print].

- Probst, A. J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J. B., Rattei, T., et al. (2014). Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat. Commun.* 5, 5497. doi: 10.1038/ncomms6497
- Qian, J., Garrett, T. A., and Raetz, C. R. H. (2014). In vitro assembly of the outer core of the lipopolysaccharide from *Escherichia coli* K-12 and *Salmonella typhimurium*. *Biochemistry* 53, 1250–1262. doi: 10.1021/bi4015665
- Raaijmakers, J. M., de Bruijn, I., and de Kock, M. J. D. (2006). Cyclic lipopeptide production by plant-associated *Pseudomonas* spp.: diversity, activity, biosynthesis, and regulation. *Mol. Plant Microbe Interact.* 19, 699–710. doi: 10.1094/MPMI-19-0699
- Raymann, K., Forterre, P., Brochier-Armanet, C., and Gribaldo, S. (2014). Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea. *Genome Biol. Evol.* 6, 192–212. doi: 10.1093/gbe/evu004
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352
- Rodriguez-R, L. M., and Konstantinidis, K. T. (2014). Bypassing cultivation to identify bacterial species. *Microbe* 9, 111–118.
- Rudolph, C., Moissl, C., Henneberger, R., and Huber, R. (2004). Ecology and microbial structures of archaeal/bacterial strings-of-pearls communities and archaeal relatives thriving in cold sulfidic springs. *FEMS Microbiol. Ecol.* 50, 1–11. doi: 10.1016/j.femsec.2004.05.006
- Rudolph, C., Wanner, G., and Huber, R. (2001). Natural communities of novel archaea and bacteria growing in cold sulfurous springs with a string-of-pearls-like morphology. *Appl. Environ. Microbiol.* 67, 2336–2344. doi: 10.1128/AEM.67.5.2336-2344.2001
- Safro, M., Moor, N., and Lavrik, O. (2004). “Phenylalanyl-tRNA Synthetases,” in *The Aminoacyl-tRNA Synthetases*, eds M. Ibba, C. S. Francklyn, and S. Cusack (Geogre Town, TX: Landes Biosciences), 250–265.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

- Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., and Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* 10, 1696–1705. doi: 10.1038/ismej.2015.233
- Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112
- Snel, B., Bork, P., and Huynen, M. (2000). Genome evolution. Gene fusion versus gene fission. *Trends Genet.* 16, 9–11. doi: 10.1016/S0168-9525(99)01924-1
- Stahl, D. A., and Amann, R. (1991). “Development and application of nucleic acid probes,” in *Nucleic Acid Techniques in Bacterial Systematics*, eds E. Stackebrandt and M. Goodfellow (New York, NY: John Wiley & Sons), 205–248.
- Suhre, K., and Claverie, J.-M. (2004). FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.* 32, D273–D276. doi: 10.1093/nar/gkh053
- Suzuki, T., Iwasaki, T., Uzawa, T., Hara, K., Nemoto, N., Kon, T., et al. (2002). *Sulfolobus tokodaii* sp. nov. (f. *Sulfolobus* sp. strain 7), a new member of the genus *Sulfolobus* isolated from Beppu Hot Springs, Japan. *Extremophiles* 6, 39–44. doi: 10.1007/s007920100221
- Swan, B. K., Chaffin, M. D., Martinez-Garcia, M., Morrison, H. G., Field, E. K., Poulton, N. J., et al. (2014). Genomic and metabolic diversity of marine group i thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS ONE* 9:e95380. doi: 10.1371/journal.pone.0095380
- Takai, K., and Horikoshi, K. (2000). Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl. Environ. Microbiol.* 66, 5066–5072. doi: 10.1128/AEM.66.11.5066-072.2000
- Teske, A., Hinrichs, K.-U., Edgcomb, V., Gomez, A., de, V., Kysela, D., et al. (2002). Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Appl. Environ. Microbiol.* 68, 1994–2007. doi: 10.1128/AEM.68.4.1994-2007.2002

- Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., et al. (2003). The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. U.S.A.* 100, 12984–12988. doi: 10.1073/pnas.1735403100
- Westram, R., Bader, K., Pruesse, E., Kumar, Y., Meier, H., Glöckner, F. O., et al. (2011). “ARB: a software environment for sequence data,” in *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, ed. F. J. de Bruijn (Hoboken, NJ: John Wiley & Sons), 399–406.
- Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337, 1661–1665. doi: 10.1126/science.1224041
- Wu, M.-C., Lin, T.-L., Hsieh, P.-F., Yang, H.-C., and Wang, J.-T. (2011). Isolation of genes involved in biofilm formation of a *Klebsiella pneumoniae* strain causing pyogenic liver abscess. *PLoS ONE* 6:e23500. doi: 10.1371/journal.pone.0023500
- Yu, Y., Lee, C., and Hwang, S. (2005). Analysis of community structures in anaerobic processes using a quantitative real-time PCR method. *Water Sci. Technol.* 52, 85–91.
- Zhang, K., Martiny, A. C., Reppas, N. B., Barry, K. W., Malek, J., Chisholm, S. W., et al. (2006). Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* 24, 680–686. doi: 10.1038/nbt1214
- Zillig, W., Stetter, K. O., Wunderl, S., Schulz, W., Priess, H., and Scholz, I. (1980). The *Sulfolobus*-“*Caldariella*” group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. *Arch. Microbiol.* 125, 259–269. doi: 10.1007/BF00446886

Chapter III: Atribacteria support community-wide microbial persistence through 44,000 years of Baltic Sea sedimentation

This chapter has been prepared for submission as a peer-reviewed article in Science authored by Jordan T. Bird, Eric Tague, Laura Zinke, Jenna M. Schmidt, Andrew D. Steen, Brandi Reese, Ian P. Marshall, Gordon Webster, Andrew Weightman, Hector Castro, Shawn Campagna, and Karen G. Lloyd.

Jordan T. Bird, Karen G. Lloyd, Hector Castro, Brandi Reese, and Andrew Steen designed the study. All analyses were completed by Jordan T. Bird, Eric Tague, Laura Zinke, and Jenna M. Schmidt. Other genomic data was provided by Brandi Reese, Ian P. Marshall, Gordon Webster, and Andrew Weightman. All authors contributed to the discussion of the results. The manuscript was written by Jordan T. Bird with inputs from Karen G. Lloyd, Eric Tague, and Laura Zinke.

Abstract

Energy-starved microbes in deep marine sediments are active and abundant, persisting in near-zero growth states for thousands of years with unknown mechanisms. In this study, we investigate deep Baltic Sea sediments with single cell genomics, metabolomics, metatranscriptomics, and enzyme assays to identify the persistence mechanisms employed by Atribacteria, Aminicenantes, Actinobacteria OPB41, Aerophobetes, Chloroflexi, Deltaproteobacteria Desulfatiglans sp., Bathyarchaeota, and Euryarchaeota Marine Group II lineages. These data indicate that transcriptional/translational regulatory elements (toxin-antitoxin systems, DNA-binding helix-turn-helix motifs, and transposases) and protective metabolite trehalose synthesis are common across lineages. Enzymatic and transcript data suggest Atribacteria and Actinobacteria OPB41 catabolize sugars and Aminicenantes and Atribacteria degrade proteins. Atribacteria uniquely access sugars, organohalides, and allantoin; utilize energy-efficient sodium pumps; and produce (and potentially export) all twenty amino acids. I conclude that growth-modulating mechanisms including regulatory elements and chemical protectants are common features in subsurface communities, and Atribacteria's specific adaptations have led to its broad success in marine sediment and its possible role as a subsurface keystone species.

Main Text

Collectively, marine deep subsurface sediments are Earth's largest reservoir of recent carbon (1), and one of its largest biospheres (2). Little is known about these ecosystems that persist in a long-term maintenance state, largely because most microbial phyla in deep subsurface marine sediments have never been cultured (5). In marine deep subsurface sediments, microbial biomass turnover times are slower than all known cultures (3), and microbes survive on less energy than growing cultures require by orders of magnitude (4). Success in an ecosystem that remains undisturbed for tens of thousands of years is likely to be determined not by the rate or efficiency of growth, but rather by the ability to decrease decay rate (6). I use multiple lines of evidence to identify mechanisms used by the *in situ* communities to gain energy, decrease decay, and interact with each other in long-term energy-starved deep subsurface sediments.

I obtained profiles of small organic molecules in deep subsurface sediment cores from IODP Expedition 347: Baltic Sea Paleoenvironment. Sampling down to 85 meters sediment depth, the expedition brought back cores that represent > 44,000 years of sedimentation (7). Forty-five single-cell amplified genomes (SAGs) from two sites, Lille Belt (M0059) and Anholt Basin (M0060), and metatranscriptomes from three sites, M0059, M0060, and Landsort Deep (M0063) were also collected (8) (Fig. 3.1, Table 3.1). Due to low transcript recovery from M0060, only the SAGs and metabolites from this site are considered in this study. The SAGs gathered from M0059 and M0060 include genomes from the three dominant bacterial phyla in 16S rRNA gene libraries of Baltic Sea sediment (Fig. 3.1, Fig. 3.2): the uncultured phylum Aminicenantes (formerly OP8), Atribacteria (formerly JS1/OP9), and Aerophobetes (formerly NT-B2). They also include genomes from uncultured groups within phyla that have cultured members: OPB41 within the Actinobacteria, relatives of *Desulfatigans sp.* within the Deltaproteobacteria, and multiple groups within the Chloroflexi (Fig. 3.1). Despite reported low abundance of archaeal cells, SAGs of Bathyarchaeota (formerly MCG) and Marine Group II in the Euryarchaeota were also recovered (9).

All the SAGs recruit transcripts, suggesting that they all represent active communities, with significantly more read recruitment among the Atribacteria, Aminicenantes, and Actinobacteria OPB41 lineages in M0059 and Atribacteria, Aerophobetes, and Aminicenantes in M0063 (Fig.

3.1B). Since each of the SAGs is incomplete (Fig. 3.1C), all the SAGs within a lineage are considered in concert. Relative to pure culture studies (9), few metabolites are identified (Fig. 3.3). The low yield is likely due to low cell abundance and activity, rather than interference from the sediment matrix during extraction, since extracting cells from sediment do not increase the number of identifiable metabolites even when the sample volume is increased by 50x. A negative control of extraction materials produces no metabolites. The metabolite extraction and detection methods do not distinguish between intra- and extra-cellular metabolites and are biased against metabolites with >100 m/z, as well as, hydrophobic, phosphorylated, and non-polar metabolites (10). Therefore, we only consider metabolites to be indicators of metabolic activity when they are also supported by genomic and transcriptomic data.

The chemical protectant trehalose is detected at nearly every depth (Fig. 3.3). In M0059, a marked increase ($\sim 10x$) in trehalose concentration is seen from 48.3 to 68.02 mbsf with a dramatic increase ($\sim 100x$) in the deepest sample. In M0060, an increase in trehalose concentration is observed from 14.3 – 40.65 mbsf, but is relatively constant with depth otherwise. Increasing trehalose with depth may indicate accumulation with sediment age or microbial responses to changes in salinity over those depths. Trehalose has an exceptional ability to stabilize proteins and osmoregulate, and it is associated with increased lifespan and decreased growth rate in eukaryotes and bacteria (9, 11). Six of the seven bacterial phyla represented by SAGs have genes encoding trehalose synthase. Four bacterial phyla encode SugA and SugB transporters and five bacterial phyla encode MalG/MalF transporters, which are known to be involved in trehalose transport. Atribacteria and, to a lesser extent, Actinobacteria OPB41 and Chloroflexi, recruit mRNA reads to trehalose synthase and trehalose transporters (Supplemental Data). Trehalose synthase is rare among genomes from cultured organisms (2.9% of archaea and 0.1% of bacteria, searching for genes annotated as EC 2.4.1.245 on JGI/IMG). Therefore, the ability to synthesize and/or import the protective disaccharide trehalose may be a general adaptation to this environment, possibly contributing to slow cell turnover rates.

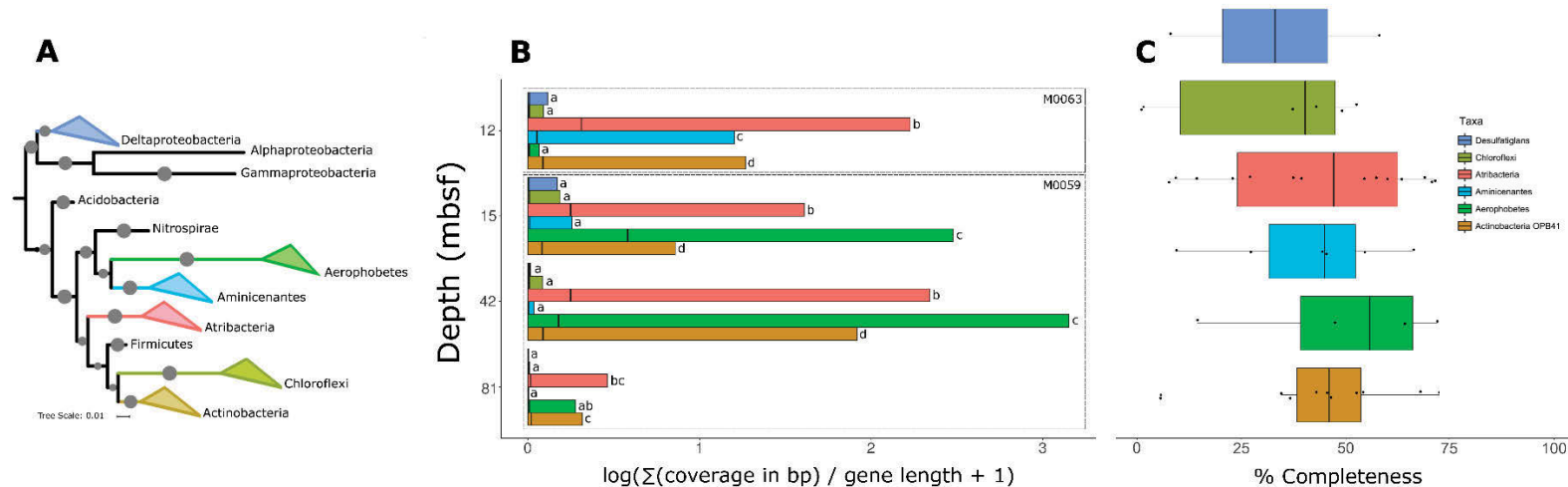


Figure 3.1: Single amplified genomes from diverse and abundant bacterial lineages. (A) A maximum-likelihood phylogenetic tree was constructed using full length 16S rRNA genes derived from sequenced SAGs and representative bacterial sequences from SILVA v123. Gray dots are proportional to bootstrapped support (80% - 100%) and colored triangle represent SAGs lineages with members sequenced in this study. (B) Colored box and whisker plots represent the range of transcript recruitment from the indicated transcriptomes to individual genes within lineage member). The mean transcript recruitment for each lineage is indicated by the black bar, while the edges of the box are defined by the 1st and 99th percentile values for read recruitment. (C) Colored box and whisker plot displayed the range, median, and quartiles of the calculated genome completeness for each SAG lineage, while staggered dots indicate the individual completeness values.

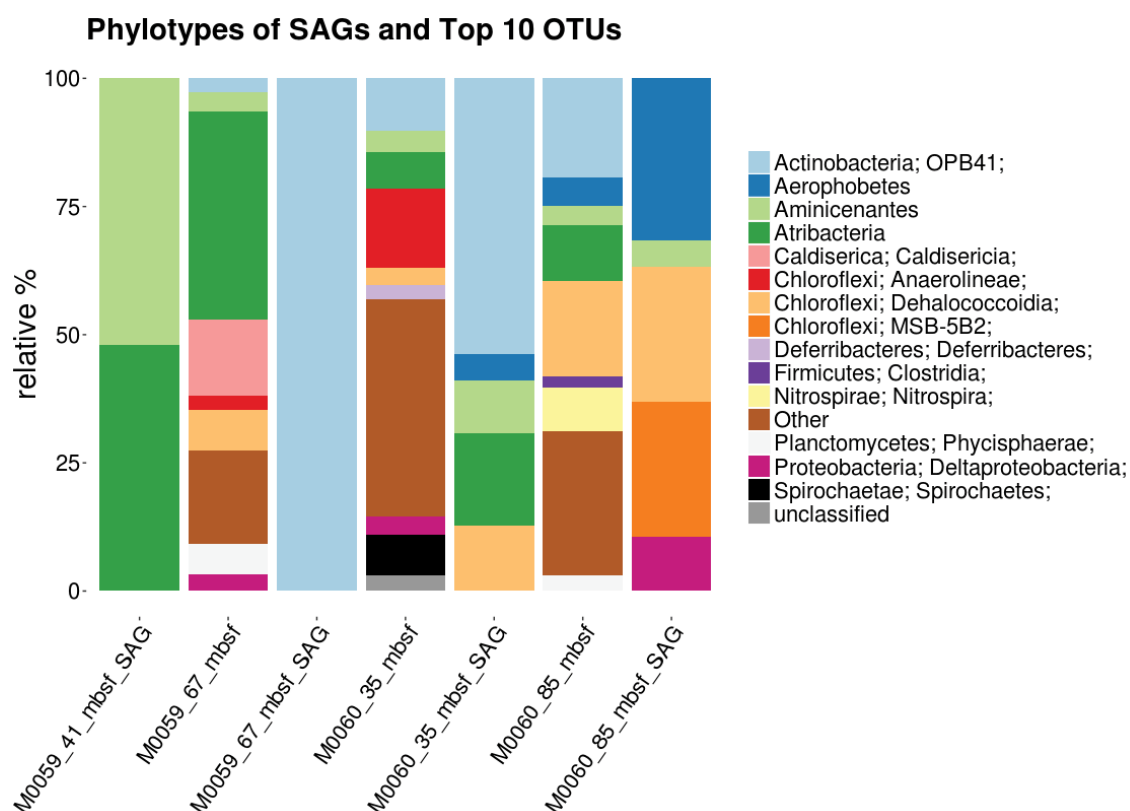


Figure 3.2: Operational taxonomic units (OTUs) composition for three 16S rRNA gene based microbiomes of Baltic sea sediment horizons are displayed in the stacked bar graphs. The taxonomy of each of the top 10 most abundant OTUs is detailed based on its closest match in the SILVA 119 database with some corrections for recently named taxonomies. The label “Other” represents the proportion of OTUs not within the top ten in abundance. The taxonomy and composition of the SAGs recovered are represented in the stacked bar graphs with the “SAG” label.

Table 3.1: Genome sources and accessions

Genome	Platform	Read Format	Location	Accession	Completeness	Redundancy
Chl_A19	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.3741	0
Chl_A21	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.4892	0
Chl_B06	HiSeq	2x125	MBL, Woods Hole, MA	PRJNA417388	0.0144	0
Chl_C14	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.5252	0
Chl_G11	HiSeq	2x125	MBL, Woods Hole, MA	PRJNA417388	0.0144	0
Chl_N02	HiSeq	2x125	MBL, Woods Hole, MA	PRJNA417388	0.4317	0
Dsu_D02	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.0791	0.0144
Dsu_E09	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.5827	0.0144
JS1_A09	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.3957	0
JS1_C14	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.1871	0
JS1_D03	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.7122	0.4964
JS1_E13	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.6906	0.5827
JS1_E15	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.0791	0
JS1_E20	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.3741	0.0144
JS1_F07	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.2734	0
JS1_I07	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.2302	0.0216
JS1_K04	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.5755	0.1439
JS1_L14	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.0935	0.0072
JS1_L23	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.1439	0
JS1_M10	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.705	0.0288
JS1_M21	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.5468	0.0216
JS1_N06	MiSeq	2x300	University of Edinburgh, UK	PRJNA417388	0.5971	0.5108
JS1_O21	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.6331	0
MCG_K23	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.358	0
MG2_P15	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.1727	0
NT-A5	MiSeq	2x250	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.4748	0.0072
NT-E05	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.7194	0.0144
NT-P03	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.1439	0
NT-P19	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.6403	0
OP8_C16	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.6619	0
OP8_E21	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.446	0.0072
OP8_F13	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.5468	0.0072
OP8_M19	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.2734	0.0072

Table 3.1 Continued: Genome sources and accessions

Genome	Platform	Read Format	Location	Accession	Completeness	Redundancy
OP8_M21	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.4532	0
OP8_P22	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.0935	0
OPBA10	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.6763	0.0144
OPBB05	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.4676	0.0072
OPBB07	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.4317	0.0144
OPBC09	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.7266	0.0144
OPBI09	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.3453	0.0144
OPBM06	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.0576	0
OPBM19	HiSeq	2x125	MR DNA Laboratory, Shallowater, TX	PRJNA417388	0.3669	0
OPBM23	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.5396	0
OPBO21	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.5252	0.0144
OPBO22	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.4532	0
Unk_M15	MiSeq	2x250	University of TN, Knoxville, TN	PRJNA417388	0.0647	0

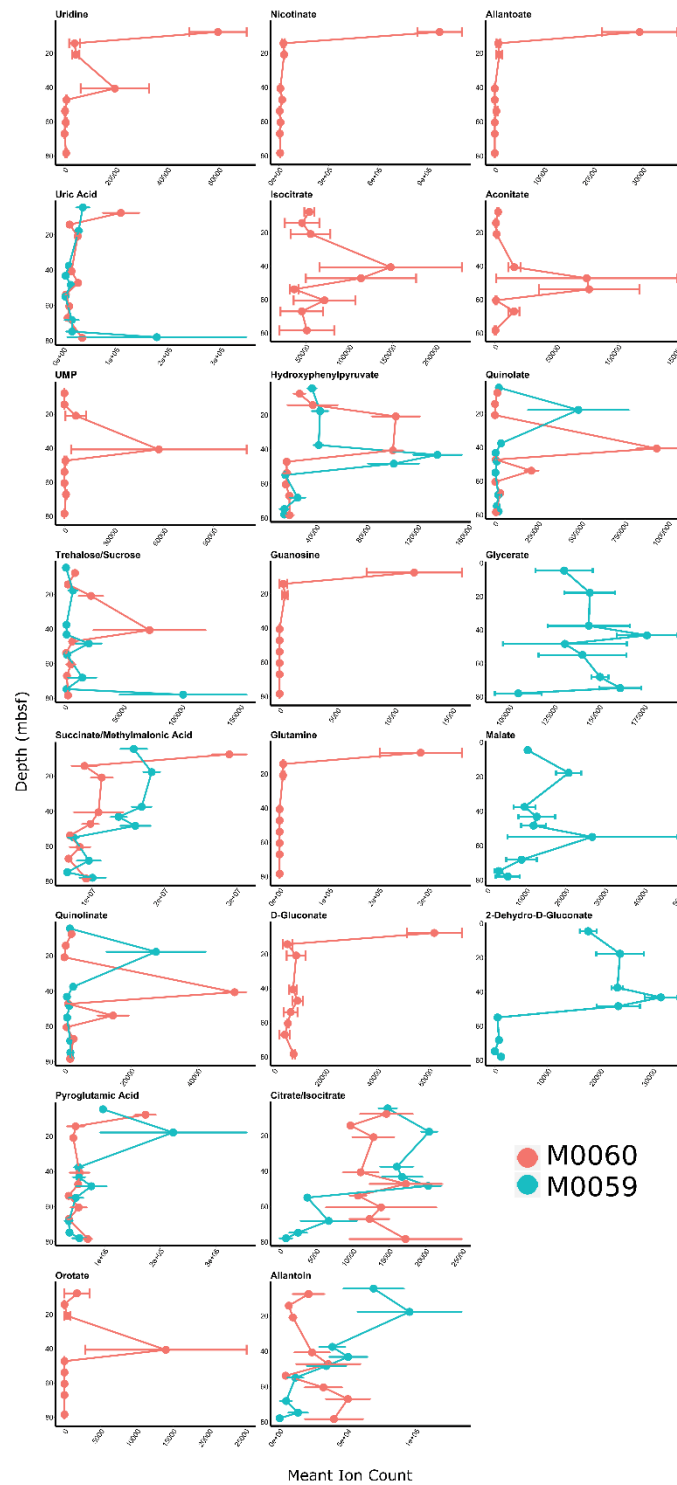


Figure 3.3: Downcore profiles of each small organic molecule detected in the Baltic Sea sediment. Error bars represent standard error from triplicate measurements.

Each of these subsurface lineages have regulatory elements in their genomes which were also highly expressed (Fig. 3.4). One of the only model systems for long-term persistence is *Mycobacterium tuberculosis*. In *M. tuberculosis*, the global transcription regulators SigE and SigH determine entrance and exit from a persistence state (12). Regulatory proteins often have helix-turn-helix motifs for DNA binding (13). Bacterial lineages in this study average between 21 and 38 proteins per million basepairs in helix-turn-helix domains, including homologs to SigE and SigH (Fig. 3.4A). Atribacteria actively transcribe nearly all of these potential transcriptional regulators in the three marine transcriptomes (Fig. 3.4B). *Escherichia coli* use the type II toxin-antitoxin systems, MazEF and RelBE, to reduce translation under unfavorable growth conditions, especially under amino acid starvation (14). Toxin-antitoxin systems are also thought to play a role in microbial persistence of slow-growing subpopulations that are resistant to antibiotics and other environmental stressors (15). Over 300 genes across all 45 SAGs encode toxin-antitoxin systems with Type II toxin-antitoxin systems being the most common among them. After accounting for the incompleteness of the SAGs, we predict an average of 30 toxin-antitoxin (Type II and IV) systems per SAG, which is similar to the complement found in *M. tuberculosis*, and far more than the average of eight toxin-antitoxin systems found per genome in all cultured organisms (16). These data suggest that global transcriptional remodeling allows long-term microbial persistence in the deep subsurface, perhaps by limiting energetic expenditures on translation of unneeded proteins, which has been predicted to be the most expensive but unavoidable expenditure in energy-starved populations (17). The transcriptomic emphasis on gene regulation may also facilitate metabolic flexibility within single cells without requiring the growth of subpopulations with genotypic variants, which would require more energy.

We observe consistent high-level expression of various transposases across the most abundant lineages (Fig. 3.4). Transposase activity may contribute to the genome plasticity we see across genomes of the same lineage and facilitate phenotypic changes under conditions where mutations due to replication error have been shown to be rare (6).

Gene-only metabolic reconstructions for many of these lineages predict extracellular hydrolysis of peptides and carbohydrates (18, 19). Extracellular Atribacteria and Actinobacteria OPB41 transcript abundance correlates with carbohydrate hydrolase activity (α -glucosidase, β -glucosidase, and N-acetyl- β -glucosaminidase) (Fig. 3.5), with Atribacteria dominating transcript

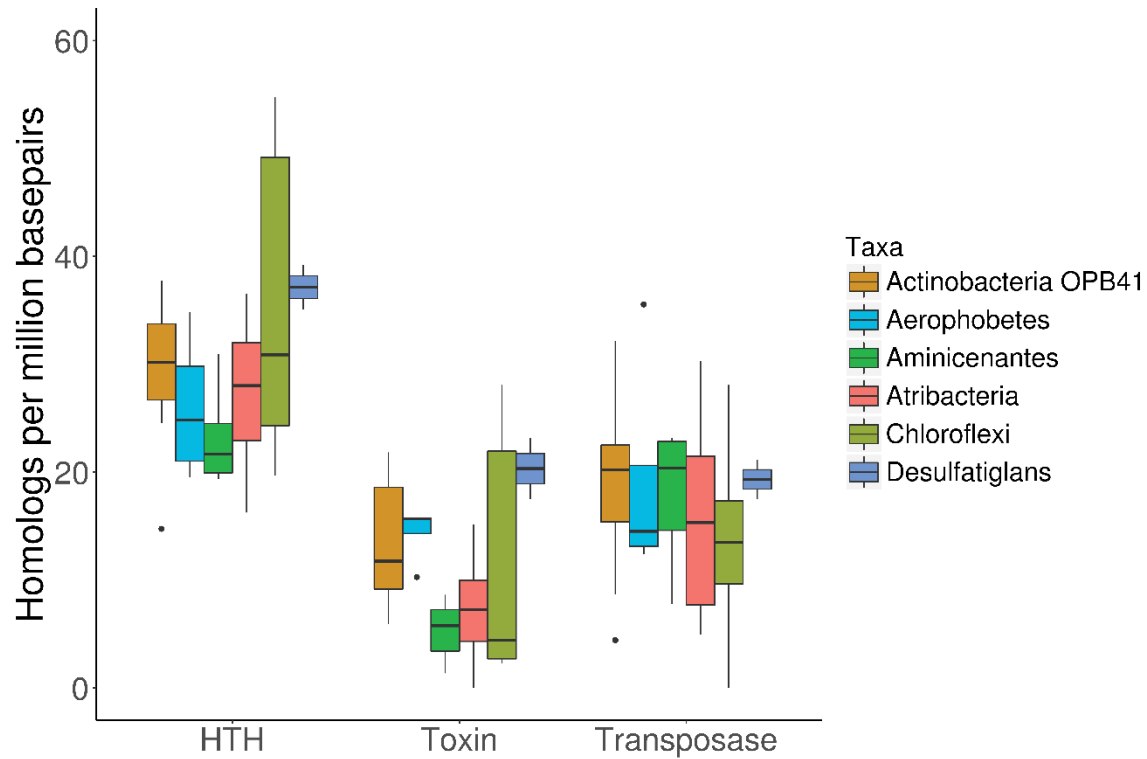


Figure 3.4A: Transcriptional and translational regulatory elements are abundant and expressed. (A) Colored box and whisker plots display the prevalence of helix-turn-helix (HTH) DNA binding domains, Type II and Type IV toxin/antitoxin proteins, and genes containing domains indicative of transposases with each SAG normalized to a million basepairs. Chl_B06 contained 112.3 HTH domains per million pairs and was not displayed in this graph.

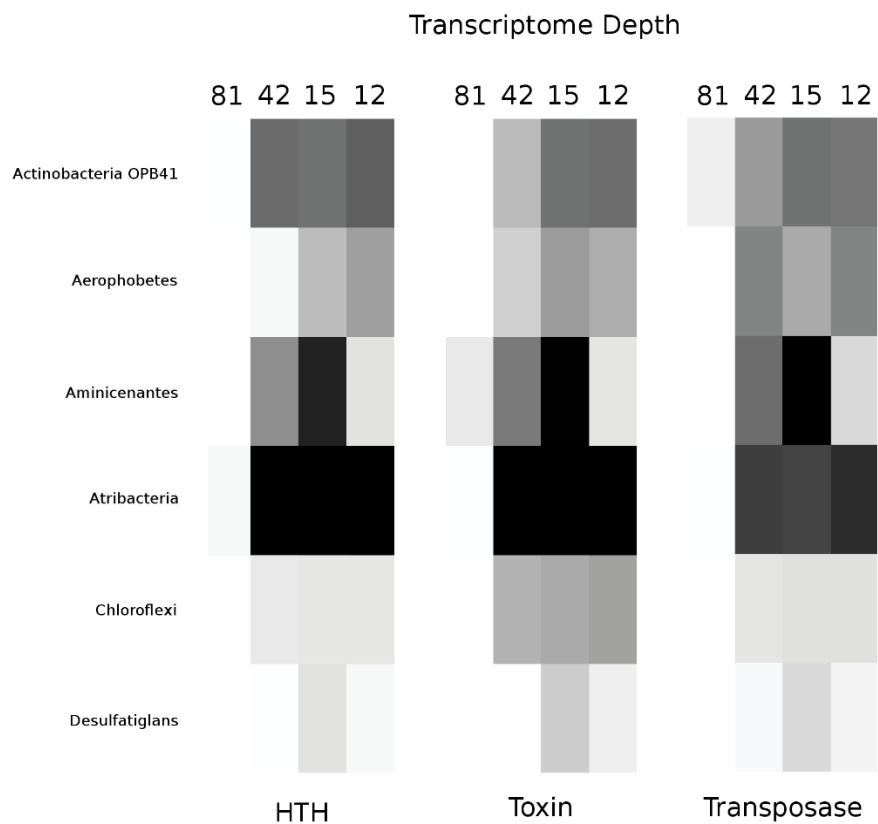


Figure 3.4B: Transcriptional and translational regulatory elements are abundant and expressed. The heatmaps indicate the proportions (white to black, 0 to 100%) of the regulatory elements in Figure 3.4A that were also detected in the metatranscriptomes.

abundance. Carbohydrate-active enzymes were some of the most highly expressed genes in these clades and in *Aerophobetes*, but the latter expresses galactosidases whose activity was not measured. Collectively, these data support catabolic niche partitioning with specific lineages expressing hydrolases for distinct classes of the macromolecules within the shared detrital carbon pool.

In agreement with previous studies, transcript recruitment is consistently high for Atribacteria SAGs at all depths (Fig 3.1B) indicating that the group is especially successful in organic rich marine sediments (21). We identify mechanisms unique to Atribacteria that could account for their success. Atribacteria SAGs encode allantoinase, which degrades allantoin, an N-rich organic monomer (Fig. 3.5). Allantoin, a degradation product of purine catabolism, was present at all depths in M0060 and M0059 (Fig 3.3). Additionally, metabolites upstream of allantoin in purine degradation, guanosine and uric acid, and a downstream degradation product, allantoate, are in the topmost sample of M0060 (Fig 3.3, Fig 3.5). Allantoin degradation genes (ATP:ammonium generating kinase, carbamoyltransferase, and allantoinase) appear to be in operons since they are adjacent to each other on contigs, in the same reading frames, and recruit similar amounts of transcripts (Fig 3.5). One allantoinase operon has an upstream transcription regulatory element from the GntR family, which represses allantoin degradation in *Klebsiella pneumoniae* in aerobic conditions (22). This operon has high transcript abundance for GntR and the lowest transcript abundance for allantoinase. Atribacteria also contains three possible transporters for allantoin. In two Atribacteria SAGs, an ATP dependent transporter with a purine binding protein encodes directly behind allantoin degradation genes, but does not recruit transcripts across the length of the gene. Instead, a nearby ATP-independent TRAP-type C4-canters transcripts for allantoinases.

In addition to being an unique catabolic substrate to the Atribacteria lineage, allantoin degradation provides, carbamoyl phosphate, a necessary anabolic substrate for proline, arginine, and nucleotide synthesis. Allantoin degradation provides the sole route for carbomoyl phosphate production, since the common L-glutamine pathway (24) is absent in all 15 Atribacteria SAGs. Transcript recruitment to the downstream pathways for UTP and CTP synthesis and the detection of orate and UMP metabolites suggest this pathway is active (Fig. 3.3, Fig. 3.6). Exogenously-administered allantoin increases the lifespan of *Caenorhabditis elegans* with similar effects to

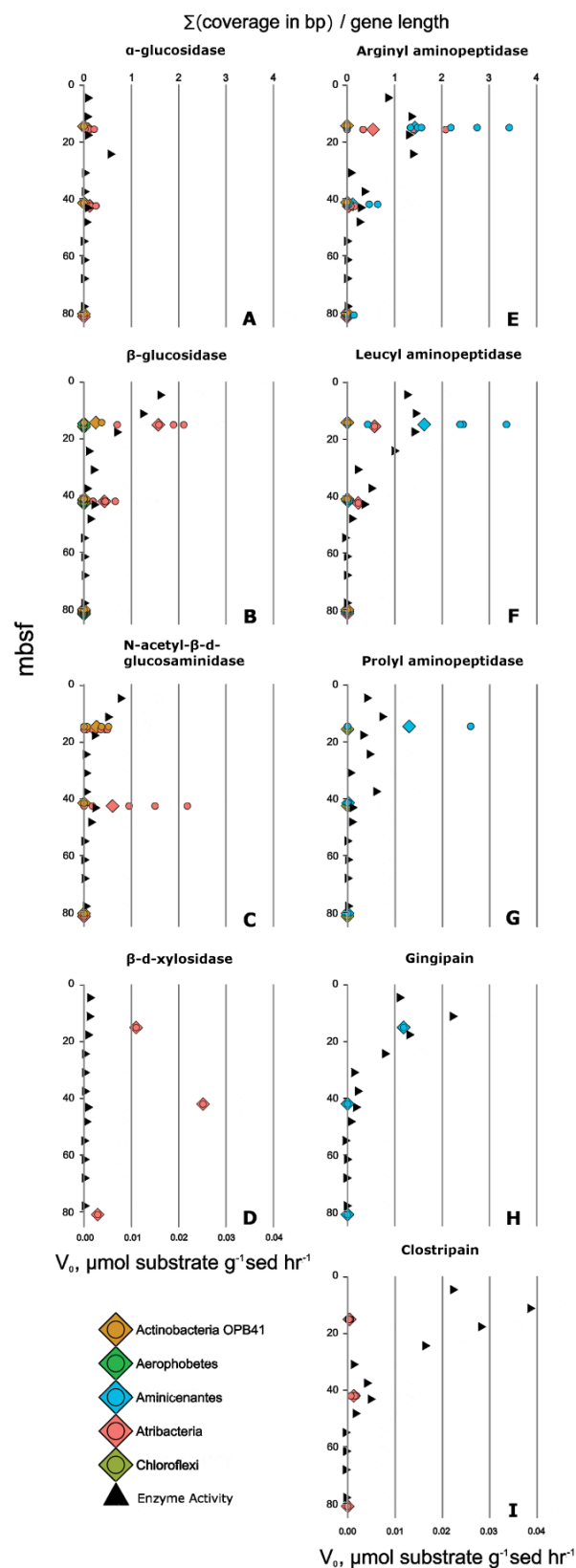


Figure 3.5: Correlations between enzyme activities and mapped transcript abundances for genes containing the appropriate functional domain (circles) and the mean abundance across each lineage (diamonds) in M0059. Expression levels for genes were plotted alongside *in-situ* activity measurements in the following protein domain annotations and substance proxies combinations: PS01324 and MUB-α-D-glucopyranoside (A), TIGR03356 and MUB-β-D-glucopyranoside (B), PTHR30480 and MUB-N-acetyl-β-D-glucosaminide (C), PF04616 and MUB-β-D-xylopyranoside (D), PF03577 and L-arginine-AMC (E), PTHR12147 and Leucine-AMC (F), PTHR10804:SF17 and H-proline-AMC (G), PF01364 and Z-phenylalanine-arginine-AMC (H), PF03415 and Z-phenylalanine-valine-arginine-AMC (I).

caloric restriction (i.e. broad downregulation in transcription) (25), suggesting that allantoin may also serve a dual role similar to trehalose by promoting slow growth. Atribacteria also appear to be uniquely positioned to catabolize organohalides, with high transcript abundances for five reductive dehalogenases. None of these, however, are associated with a membrane complex as they lacked the necessary signaling motifs to target the enzyme to the membrane (26). However, these data do suggest a catabolic role for these dehalogenases allowing Atribacteria to access another recalcitrant pool of carbon.

Atribacteria genomes in this study express genes for an energy-yielding respiration complex. Atribacteria are the only bacteria in this study that encode and express a membrane-bound hydrogenase complex with an adjacent oxidoreductase that might use sulfur compounds (27). This complex may accept electrons from reduced ferredoxin or possibly NADH, and it either produces H_2 from two protons and/or reduces sulfur compounds to create a proton gradient across the membrane. This hydrogenase complex is flanked by two multi-gene Mrp systems, Na^+/H^+ antiporters common in organisms that rely on a Na^+ motive force. Atribacterial ATP synthases were specific for Na^+ rather than H^+ , and were expressed alongside pyrophosphatase/ Na^+ pumps which may help maintain the Na^+ gradient by utilizing energy from pyrophosphate that would otherwise go to waste (Fig. 3.6). Organisms under energy-limitation tend to utilize Na^+ rather H^+ motive force, since cell membranes are less permeable to Na^+ , and H^+ gradients can be dissipated by extracellular organic acids (28). We propose that the great success of Atribacteria in organic-rich marine sediments (21) is due to its ability to catabolize a wider range of substrates than other members of the community including allantoin, many carbohydrates, some peptides, and organohalides. This catabolism results in substrate level phosphorylation as well as reduced cofactors which produce H_2 , producing a H^+ motive force which is converted to a Na^+ motive force by antiporters to maximize energetic efficiency.

Atribacteria utilize at least some of this energy to meet their highest energy demand, repairing proteins (17). The effects of energy stress on proteins are evinced by the presences of Pyroglutamic acid, a degradation product of glutamate that accumulates with age in peptides *in vitro* (29), at both M0060 and M0059. Atribacteria is the only lineage that contained all the genes for *de novo* synthesis of amino acids, all of which were transcribed. Hydroxyphenylpyruvate, a key intermediate in the synthesis of aromatic amino acids and a direct

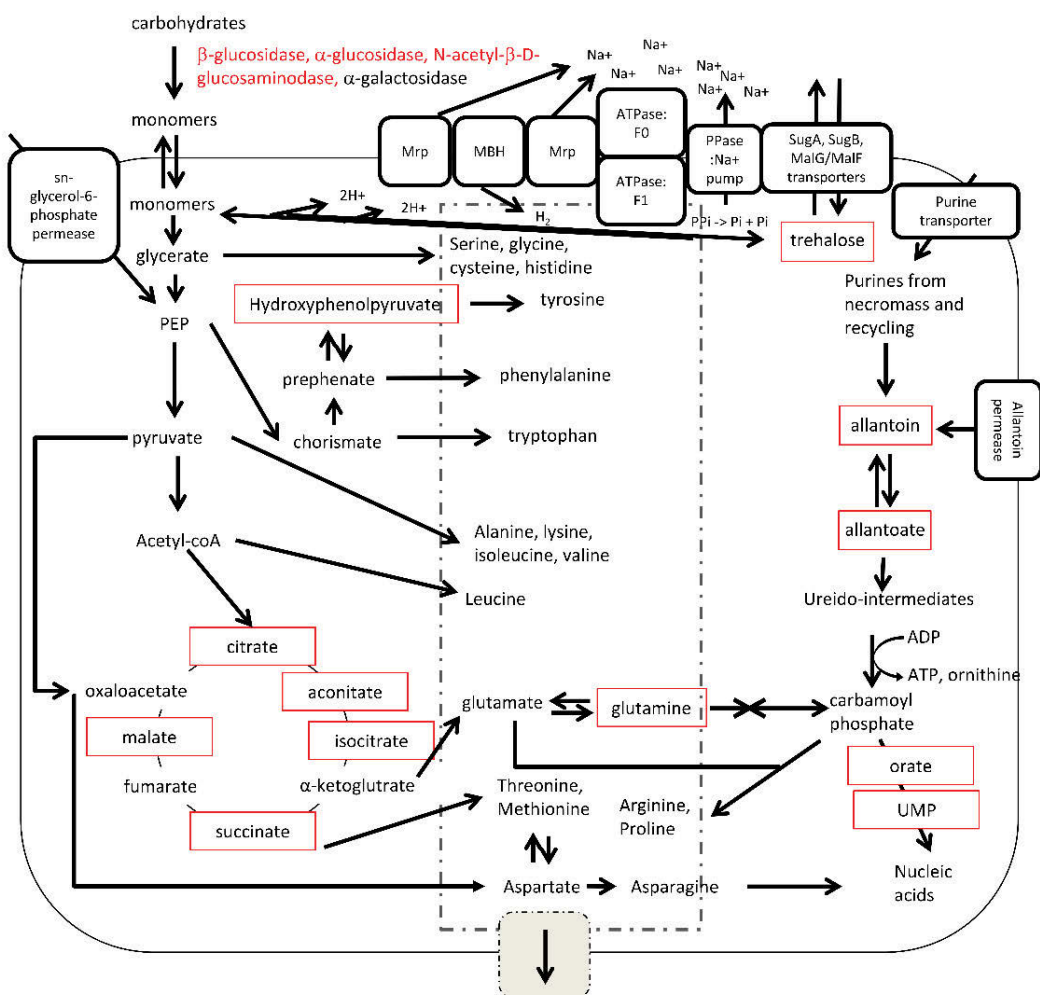


Figure 3.6: Unique attributes of Atribacteria physiology provide an advantage in energy-limited marine sediment systems. Key metabolites in metabolic pathway were written out with those detected in small-organic molecule profiles in red boxes. Red text is used for enzymes for which activity measurements were taken (see Fig. 3.4). Key enzymes and transporter complexes are represented as black boxes. Black arrows indicate enzymatic processes or pathways for which transcripts black boxes. Black arrows indicate enzymatic processes or pathways with transcripts that mapped to the Atribacteria lineage. The dotted box around the amino acids highlights the complete suite of amino acids produced by Atribacteria which could be exported from the cell via an encoded amino acid export pump.

precursor to tyrosine, is present in both sites down to 55 mbsf. Its abundance positively correlates with overall Atribacteria transcript recruitment (Fig 3.3, Fig. 3.6). Atribacteria also recruit transcripts for general amino acid exporters, including the highly expressed aromatic amino acid exporter YddG (Fig. 3.5, Supplemental Data Table 3.1). However, only one free amino acid, glutamine, is among our metabolites (Fig. 3.3). This amino acid appeared only in M0060 and decreased rapidly with depth, disappearing completely by 41 mbsf. The absence of many amino acids in the metabolite pool could be due to uptake from other community members, such as Actinobacteria-OPB41, Aerophobetes, and *Desulfatiglans sp.*, which transcribe genes for amino acid uptake enzymes (Supplementary Data Table 3.1). The unique ability of Atribacteria to synthesize and export amino acids *de novo* in this energy-limited environment points to their potential role as a keystone species, cross-feeding amino acids to other starving microbes. Amino acid production and export meet Atribacteria's two competing needs: repairing proteins while maintaining a near-zero growth rate. Cellular growth rate positively correlates with the intracellular concentration of amino acids, and below a threshold value, growth ceases entirely (30). Atribacteria benefit from the amino acid sink provided by other community members, since this allows amino acid export to occur in the direction of the gradient.

By linking *in situ* metabolites with SAGs, transcripts, and enzyme activities, we found that energy-limited deep subsurface microbes utilize the chemical protectant trehalose, as well as, type II and type IV toxin-antitoxin systems and other regulatory elements that may inhibit growth rates and tightly regulate expensive protein production. Correlations between measured enzyme activities and specific enzyme transcripts suggest catabolic niche partitioning between subsurface community members based on their ability access specific carbon macromolecules. Atribacteria perform catabolic and anabolic functions, including the degradation of allantoin and organohalides and the *de novo* synthesis and export of amino acids, which may help Atribacteria and the rest of the community persist over long-term burial in saline, anoxic sediment environments.

Materials and Methods

Metabolite Extraction from Marine Sediments

Sediment samples were removed from a -80°C freezer and within five minutes shavings were transferred to sterile plastic tubes in 100 mg triplicate subsamples and returned to the freezer. Samples were later ground to a powder using a mortar and pestle sitting in liquid nitrogen. Technical replicates were created by portioning out approximately 100 mg of ground sample. 1.3 mL of chilled extraction solvent (40:40:20, acetonitrile: methanol: water +0.1M formic acid) were added and the mixture was vortexed to suspend the sample. The extraction process lasted 20 minutes at 4°C on an orbital shaker, and then the samples were centrifuged at 16,800 x g for 5 minutes. The supernatant was removed and collected in a separate vial. Then, a second extraction was performed using a 200 µL of extraction solvent and the same extraction process. After both volumes of supernatant were collected in the same vial, the solvent was evaporated to dryness using nitrogen gas. The dried samples were then suspended using 300µL of sterile water and transferred to autosampler vials for analysis.

UPLC/MS

Mass spectrometric analysis was performed using an Exactive Plus orbitrap mass spectrometer (Thermo Scientific) coupled to a UPLC system (Dionex). The separation was achieved using a Synergi HydroRP column (Phenomenex) with a solvent gradient modeled after (31). The sample was introduced using electrospray ionization and full scan data from 85 *m/z* to 1000 *m/z* was collected in negative mode.

Data Manipulation

Data generated from the mass spectrometer's software was converted using the MSConvert package of the ProteoWizard (32) toolkit software. Files were then imported into the MAVEN (33) software to visualize chromatograms for specific *m/z* ratios. Using a list of known retention times and exact *m/z* values (± 5 ppm), metabolites were selected, and ion counts were compiled from areas under the curve.

Single-cell Extractions

Cell extraction methods followed a previously published one (20), with the following minor modifications. Samples were preserved in 1:1 (glycerol:sediment) mixture frozen on ship. The samples were shipped back to the University of Tennessee on dry ice. 3 ml of the sample

mixture were removed into a sterile 15 ml plastic tube. An additional artificial seawater buffer (ASW), either 30% or 15% depending on the sample salinity was added to bring the volume up to 10 mL. The tip of a Misonix Microson™ Ultrasonic Cell Disruptor was placed in an ice bath next to the plastic tube containing the sample and manually sonicated two times per second at 20% power to dislodge cells from their associated minerals. Next, the tube was vortexed for 10 s and the sediment was allowed to settle for 10 min. 5 mL aliquots of the supernatant were gently laid on top of an equal volume of 60% Nycodenz solution in a sterile 15 mL tubes. The 15 mL tubes were then centrifuged at $11,617 \times g$ and 4°C for 1 h in fiberlite rotor. When the sediment in the sample climbed up the sidewall past the Nycodenz boundary, the sample were again spun at 5,000 rpm in the swinging bucket rotor. The layer above the Nycodenz gradient was carefully removed and pooled into a sterile 15 ml plastic tube. Lastly, 1.5 ml aliquots were mixed with 375 µl of a 5% glycerol 1x TE solution and immediately placed at -80°C.

Transcriptomes

Approximately 7.5 g of frozen sediment was chipped from whole round cores in a clean room using flame sterilized instruments. Samples were extracted from the middle of the core which was determined to have little or no contact with the drill fluid (7). RNA was extracted from sediment using the MoBio PowerSoil RNA kit following manufacturer's instructions (MoBio Laboratories, Carlsbad, CA).

Total RNA extractions were treated with Ambion Turbo DNase (ThermoFisher Scientific, Waltham, MA) according to manufacturer protocols. Resulting RNA purity and quantity was checked using the Eppendorf Biospectrometer (Eppendorf, Hauppauge, NY) and by reverse transcription followed by PCR amplification of 16S SSU rRNA. DNase-treated RNA extract was PCR amplified to determine if DNase treatment was effective. RNA extracts were shipped on dry ice to the Molecular Research DNA Laboratory, LLC in Shallowater, Texas for library preparation and sequencing analysis.

For metatranscriptomes, complementary DNA synthesis was performed using the Nextera RNA sample preparation kit following manufacturer's instructions (Illumina, San Diego, CA). Libraries were sequenced on the Illumina HiSeq 2500 platform (Illumina, San Diego, CA) for 500 cycles with 250 bp paired-end chemistry. Transcripts were trimmed with TRIMMOMATIC using the ILLUMINACLIP:NexteraPE-PE.fa:2:30:12:1:true LEADING:3

TRAILING:3 SLIDINGWINDOW:10:20 MINLEN:75 parameters. Bowtie2 and Samtools were used to filter reads matching the PhiX internal standard. And finally, reads were mapped to each of the SAG contigs with Bowtie2 using the default settings and combined by concatenating mapping from the same sample (34).

Genome Annotation and Analysis

Sequence data from 5 separate Illumina sequencing runs were combined to make the library of single amplified genome short read data in the study. For all genome libraries sequenced at the University of Tennessee and Molecular Research DNA Laboratory the Nextera library preparation was used. Genomes sequenced at the Marine Biological Laboratory used the TruSeq V2 library preparations kit (Table 3.1). Sequences from each Illumina sequencing run were quality trimmed using the TRIMMOMATIC software using the

ILLUMINACLIP:{ADAPTER}:2:30:12:1:true LEADING:3

TRAILING:3 SLIDINGWINDOW:10:20 MINLEN:75 parameters (35). The remaining high quality paired and single reads were assembled using the SPAdes 3.6 assembler with the following parameters: spades.py -m 62 -o <OUTPUT_DIR> -sc -12 <PAIRED_READS> -s <SINGLES> -t 15 -k 21,33,55,77,99,127 -careful (36). Contigs with high sequence similar to the PhiX internal standard, those with less than 1000 bp and 5x coverage, and those with contigs containing only a single repeated nucleotide were removed. The remaining contigs from each genome were then annotated using the Prokka genome annotation software with a custom database which combined both of Bacteria and Archaea specific databases provided with the software (37). Predicted proteomes of each genome were additionally annotated with INTERPROSCAN5 (38). Later, these annotations were combined in and visualized using the Anvi'o software (39). Anvi'o functions which annotate single copy conserved genes were used to predict genome completeness and contaminations levels. The Anvi'o software profiling functions were used to analyze the read recruitment from the transcriptomes to the annotated genome features in each of the SAGs.

Enzyme activities

Enzymes were assayed according to the procedure described by Schmidt (40). All processing was performed under an atmosphere of 100% N₂ in a glove box. Subsamples of -80 °C-frozen core rounds were removed using a hand drill fitted with an EtOH-sterilized, 1-cm

interior diameter hole saw bit. Subsamples were kept frozen until the time of the assay. Each depth was assayed on separate days. For each depth, 3 g of thawed sediment was mixed in a Waring blender for 1 minute in 100 mL sterile, anoxic, 0.2 M borate-buffered saline, per the best-practice in soils enzyme assays (41). A separate slurry for each depth was autoclaved for 60 minutes on a liquid cycle to serve as a sterile control. Immediately after preparation, 960 μ L slurry was added to a 1 cm x 1 cm semimicro-style methacrylate cuvette and amended with 40 μ L of a solution of fluorogenic enzyme substrate (Table S3.1). For each depth, three replicate “live” and three replicate “sterile” cuvettes were poured. Additionally, two ten-point calibration samples were created, using 7-amido-4-methylcoumarin (AMC) or 4-methylumbelliferone plus sediment. Cuvettes were capped, mixed, and fluorescence in each cuvette was measured immediately. The exact time of fluorescence measurement was noted. Samples were incubated at 20-22 °C in the dark, and fluorescence was measured in each cuvette approximately 5 times over the course of 24 hours. The ambient temperature was monitored during the incubation, and small variations in temperature between depths did not explain variation among depths. Calibration samples were also measured at each measurement timepoint, and sample fluorescence was calibrated separately for each timepoint.

Changes in concentration of fluorophore over time were calculated using the R package *enzalyze* (42). As described in (40), several lines of evidence suggested that increases in fluorescence over time in autoclaved sediment were due to incompletely denatured (or, denatured-then-renatured) enzymes, rather than abiotic processes. Therefore, uncorrected hydrolysis rates of “live” samples are reported as v_0 . In addition to the substrates listed in Table S2, a substrate for cellulobiose was assayed, but did not have any detectable activity.

Data Visualization

Software packages in the R statistical language (43) including *ggplot2* (44) and *heatmap.2* (45) were utilized to produce the figures in this study. ANOVA and Tukey’s mean testing were used to compare transcript recruitment between each microbial lineage using the *base* stats package in R. The Microsoft Office suite was used to produce tables and diagrams, while the GIMP image software was used to edit the appearance figures

(<https://products.office.com/en-US/>, <https://www.gimp.org/>).

Partial 16S Sequencing

In total, DNA was extracted from 9 sediment horizons and two drill fluid samples the MP Biomedical FAST DNA kit. Two primer sets targeting the v4 and v5 regions of bacteria and archaea were used to amplify 16S rRNA marker gene sequences from the extraction. For this study only bacteria amplicon libraries from M0059, M0060, and M0063 were analyzed. In addition to fused Truseq adapter sequences the DNA primer sets were as follows: Bacterial v4v5 518F 5'- CCAGCAGCYGCGGTAAN-3', Bacterial v4v5 926R 5'-

CCGTCAATTCNTTTTRAGT -3' + 5'- CCGTCAATTTCTTTGAGT -3' + 5'-

CCGTCTATTCCTTTGANT -3'. Details on the buffers and DNA purification procedures can be found at <https://vammps.mbl.edu/resources/primers.php>. 16S amplicon data resulting from sequencing on the Illumina MiSeq platform was deposited at https://vammps.mbl.edu/portals/deep_carbon/. For this study, OTUs were determined using Mothur (46) with the Arb Silva 99 NR Ref 119 database as a reference. The top ten OTU were sorted and the latter modified bar plots were generated using EXPLECET (47).

Acknowledgments

We thank the captain and crew of the R/V Greatship Manisha, IODP, European Consortium for Ocean Research Drilling, and the scientific party of IODP expedition 347. Major funding was provided by NSF OCE-1431598 (KL, JB), and NSF Center for Dark Energy Biosphere Investigations (subawards 42525882 and 157595) (KL, JB), USSSP Schlanger Drilling Fellowship (LAZ), and some DNA sequencing by the Deep Carbon Observatory Census of Deep Life.

List of References

1. S. Arndt *et al.*, *Earth-Sci. Rev.* **123**, pg. 53–86 (2013).
doi:10.1016/j.earscirev.2013.02.008
2. J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, S. D'Hondt, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 40 (2012). doi:10.1073/pnas.1203849109
3. S. Braun *et al.*, *Sci. Rep.* **7**, 5680 (2017). doi:10.1038/s41598-017-05972-z
4. D. E. LaRowe, J. P. Amend, *Extreme Microbiol.*, **6**, 718 (2015).
doi:10.3389/fmicb.2015.00718
5. A. Teske, K. B. Sørensen, *ISME J.* **2**, 1 (2008). doi:10.1038/ismej.2007.90
6. P. Starnawski *et al.*, *Proc. Natl. Acad. Sci.* **114**, 11 (2017).
doi:10.1073/pnas.1614190114.
7. T. Andrén, B. B. Jørgensen, C. Cotterill, S. Green, and the Expedition 347 Scientists, Expedition 347 of the mission-specific drilling platform from and to Kiel, Germany Sites M0059–M0067 12 September–1 November 2013.
8. L. A. Zinke *et al.*, *Environ. Microbiol. Rep.* **9**, 5 (2017), doi:10.1111/1758-2229.12578.
9. M. J. Brauer *et al.*, *Proc. Natl. Acad. Sci.* **103**, 51 (2006). doi:10.1073/pnas.0609508103
10. W. Lu *et al.*, *Anal. Chem.* **82**, 8 (2010). doi:10.1021/ac902837x
11. P. Kyryakov *et al.*, *Front. Physiol.* **3**, Article 256 (2012), doi:10.3389/fphys.2012.00256.
12. P. Du, C. D. Sohaskey, L. Shi, *Front. Microbiol.* **7**, Article 1346 (2016),
doi:10.3389/fmicb.2016.01346.
13. R. G. Brennan, B. W. Matthews, *J. Biol. Chem.* **264**, 4 (1989). pmid:2644244
14. S. K. Christensen, M. Mikkelsen, K. Pedersen, K. Gerdes, *Proc. Natl. Acad. Sci.* **98**, 25 (2001). doi:10.1073/pnas.251327898
15. E. Rotem *et al.*, *Proc. Natl. Acad. Sci.* **107**, 28 (2010). doi:10.1073/pnas.1004333107
16. Y. Shao *et al.*, *Nucleic Acids Res.* **39**, (Database issue) (2011). doi:10.1093/nar/gkq908
17. C. P. Kempes *et al.*, *Front. Microbiol.* **8**, Article 31 (2017),
doi:10.3389/fmicb.2017.00031.
18. S. A. Carr, B. N. Orcutt, K. W. Mandernack, J. R. Spear, *Front. Microbiol.* **6**, Article 872 (2015), doi:10.3389/fmicb.2015.00872.

19. C. Rinke *et al.*, *Nature*. **499**, 7459 (2013). doi:10.1038/nature12352
20. K. G. Lloyd *et al.*, *Nature*. **496**, 7444 (2013). doi:10.1038/nature12033
21. B. N. Orcutt, J. B. Sylvan, N. J. Knab, K. J. Edwards, *Microbiol. Mol. Biol. Rev.* **75**, 2 (2011). doi:10.1128/MMBR.00039-10
22. K. Guzmán, J. Badia, R. Giménez, J. Aguilar, L. Baldoma, *J. Bacteriol.* **193**, 9 (2011). doi:10.1128/JB.01450-10
23. C. Mulligan, M. Fischer, G. H. Thomas, *FEMS Microbiol. Rev.* **35**, 1(2011). doi:10.1111/j.1574-6976.2010.00236.x
24. M. Mergeay, D. Gigot, J. Beckmann, N. Glansdorff, A. Piérard, *Mol. Gen. Genet. MGG*. **133**, 4 (1974). pmid:4373646
25. S. Calvert *et al.*, *Aging Cell*. **15**, 2 (2016). doi:10.1111/accel.12432
26. H. Smidt, W. M. de Vos, *Annu. Rev. Microbiol.* **58**, 1 (2004). doi:10.1146/annurev.micro.58.030603.123600
27. S. Laska, F. Lottspeich, A. Kletzin, *Microbiol.* **149**, 9 (2003). doi:10.1099/mic.0.26455-0
28. F. Mayer, V. Müller, *FEMS Microbiol. Rev.* **38**, 3 (2014). doi:10.1111/1574-6976.12043
29. D. Chelius *et al.*, *Anal. Chem.* **78**, 7 (2006). doi:10.1021/ac051827k
30. H. Eagle, K. A. Piez, M. Levy, *J. Biol. Chem.* **236**, 2039–2042 (1961). pmid: 13725477
31. W. Lu *et al.*, Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. *Anal. Chem.* **82**, 8 (2010). doi:10.1021/ac902837x
32. M. C. Chambers *et al.*, A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 10 (2012). doi:10.1038/nbt.2377
33. E. Melamud, L. Vastag, J. D. Rabinowitz, Metabolomic analysis and visualization engine for LC-MS data. *Anal. Chem.* **82**, 23 (2010). doi:10.1021/ac1021166
34. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 4 (2012). doi: 10.1038/nmeth.1923
35. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 15 (2014). 10.1093/bioinformatics/btu170
36. A. Bankevich *et al.*, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 5 (2012). 10.1089/cmb.2012.0021

37. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 14 (2014). 10.1093/bioinformatics/btu153
38. P. Jones *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* **30**, 9 (2014). 10.1093/bioinformatics/btu031
39. A. M. Eren *et al.*, Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, doi:10.7717/peerj.1319.
40. Schmidt, Jenna Marie, "Microbial Extracellular Enzymes in Marine Sediments: Methods Development and Potential Activities in the Baltic Sea Deep Biosphere." Master's Thesis, University of Tennessee, 2016.
http://trace.tennessee.edu/utk_gradthes/4072
41. C. W. Bell *et al.*, High-throughput fluorometric measurement of potential soil extracellular enzyme activities. *J. Vis. Exp.*, e50961 (2013). doi: [10.3791/50961](https://doi.org/10.3791/50961)
42. Steen, enzalyze. GitHub Repos. (2013), (available at <https://github.com/adsteen/enzalyze>)
43. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria).
44. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2009; <http://ggplot2.org>).
45. G. R. Warnes *et al.*, *gplots: Various R Programming Tools for Plotting Data* (2016).
46. P. D. Schloss *et al.*, Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009). doi:10.1128/AEM.01541-09
47. C. E. Robertson *et al.*, Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. *Bioinformatics.* **29**, 3100–3101 (2013). doi: 2050/10.1093/bioinformatics/btt526

Table S3.1: Enzyme substrates and nominal corresponding enzymes

Enzyme	Class	Substrate	EC
Leucyl aminopeptidase	Exopeptidase	Leucine-AMC	3.4.11.1
Arginyl aminopeptidase	Exopeptidase	L-arginine-AMC	3.4.11.6
Prolyl aminopeptidase	Exopeptidase	H-proline-AMC	3.4.11.5
Ornithyl aminopeptidase	Exopeptidase	Ornithine-AMC	--
Gingipain R.	Endopeptidase	Z-phenylalanine-arginine-AMC	3.4.22.37
Clostripain	Endopeptidase	Z-phenylalanine-valine-arginine-AMC	3.4.22.8
β -d-xylosidase	Polysaccharide hydrolase	MUB- β -D-xylopyranoside	3.2.1.37
β -d-cellobiohydrolase	Polysaccharide hydrolase	MUB- β -D-cellobioside	3.2.1.91
N-acetyl- β -d-glucosaminidase	Polysaccharide hydrolase	MUB-N-acetyl- β -D-glucosaminide	3.2.1.52
β -glucosidase	Polysaccharide hydrolase	MUB- β -D-glucopyranoside	3.2.1.21
α -glucosidase	Polysaccharide hydrolase	MUB- α -D-glucopyranoside	3.2.1.20
Alkaline phosphatase	Phosphatase	MUB-PO ₄	3.1.3.1

**Chapter IV: Comparative genomic analysis reveals earliest branching
Actinobacteria class, Candidatus Osiribacteria**

This chapter is being prepared for submission as a peer-reviewed article in mSystems authored by Jordan T. Bird and Karen G. Lloyd.

Jordan T. Bird and Karen G. Lloyd designed the study. All analyses were completed by Jordan T. Bird. Both authors contributed to the discussion of the results. Manuscript was written by Jordan T. Bird with inputs from Karen G. Lloyd.

Abstract

IODP Expedition 347 uncovered the first single cell genomes of the cosmopolitan, deeply-branching, and diverse actinobacterial taxa OPB41. Phylogenies constructed from both 16S rRNA genes and concatenate proteins suggest OPB41 should be reclassified as “*Candidatus Osirisbacteria*,” the earliest known branching class of Actinobacteria. *Osirisbacteria* genomes in this study include the lowest recorded G + C content among the Actinobacteria. Additionally, they lack a homolog to the error-prone DNA polymerase, DnaE2, thought to be important in the acquisition of neofunctions across the Superphylum, Terrabacteria. Pangenomic analyses revealed components of metabolisms common to in anaerobic fermentation with the ability to access, catabolize, and import diverse sugars. Expansion of secondary metabolisms (e.g. ladderane) and increased diversity of hydrolases could indicate differences between competition levels and energy requirements of *Osirisbacteria* that are environment specific. *Osirisbacteria* as a class of Actinobacteria is restricted to anaerobic environments and the unique physiology and evolutionary history of this taxon will inform future studies of the ancestral microbes which first colonized land.

Introduction

Genetic and phylogenetic analyses of samples taken from geographically widespread and diverse environments have identified OPB41 as a discreet, yet diverse, group within Actinobacteria. OPB41 refers to the 16S sequence found by Hugenholtz et al who sampled the groundwater welling up from hot springs at Yellowstone National Park, including at the namesake Obsidian Pool (OP) (1). Advances in gene sequencing technology since the mid-1990s, including metagenomics and single-cell genomics, have resulted in giant leaps forward in our understanding of deeply-branching organisms, including the description of previously unknown phyla (2–7). Much of the interest has focused on these high taxonomic level classifications (2–7), but a careful analyses of available genomic information could elucidate the classification of OPB41 within the Actinobacteria and the group’s defining functional and genomic characteristics. Although genetic data and a some limited previous phylogenetic analysis are available for OPB41 (7, 8), no species have ever been isolated in pure culture.

The Terrabacteria superphyla describes the radiation of bacteria which include the phyla Chloroflexi, Actinobacteria, Firmicutes, Deinococcus-Thermus, and Cyanobacteria (). The

expansion of the Terrabacteria superphylum coincides with the first evidence of microbial colonization of land in the rock record and members of the Terrabacteria possess adaptations to resist ultraviolet light radiation, desiccation, high salinities, and carry out oxygenic photosynthesis (e.g. Cyanobacteria) (9). Understanding the phylogenetic relationship, as well as, the biographical and physiological differences between the uncultured taxon Actinobacteria OPB41 and other members of the Terrabacteria could inform our present understanding of microbial colonization of land. Bird et al. prepared the first ten single amplified genomes (SAGs) of OPB41 sampled from deep subsurface sediment cores at Lille Belt (M0059) and Anholt Basin (M0060) on IODP Expedition 347: Baltic Sea Paleoenvironment (10). No specific data on the morphology of OPB41 has been published, so our analysis relies on what can be discerned from available genomic, transcriptomic, and geochemical information. Samples from these marine sediments were dominated by cocci shapes with rods and vibrio shapes appearing only in relatively shallow sediments of less than 10 meters below the seafloor (mbsf), which was above SAG sampling depths (10, 11). Here we investigate the commonalities between environments where OPB41 sequences are found and relate to its intragroup phylogenies. In this study, we analyzed the phylogenetic relationship between OPB41 and all known Actinobacteria classes, using 16S rRNA gene sequences and concatenated protein-coding gene trees from available single cell amplified genomes (SAGs) and metagenome assembled genomes (MAGs). Additionally, we performed pangenomic analyses to reveal similarities and differences among the available genomic information.

Results and Discussion

Phylogenetic placement of OPB41 relative to the phylum Actinobacteria

All currently available 16S rRNA gene sequences designated as OPB41 (by the Silva 125 taxonomy) break into 5 major groups within the Actinobacteria (Fig. 4.1). Four of these groups are monophyletic, while Group 5 falls within the Coriobacteriia class. 16S rRNA genes from the eight OPB41 SAGs from the Baltic Sea deep subsurface fall into two distinct clades: Groups 1 and 2 (Figure 4.1). Group 1 was recovered in single amplified genomes (SAGs) separated from deep Baltic sediments, but did not recruit transcripts from Baltic marine sediments as well as

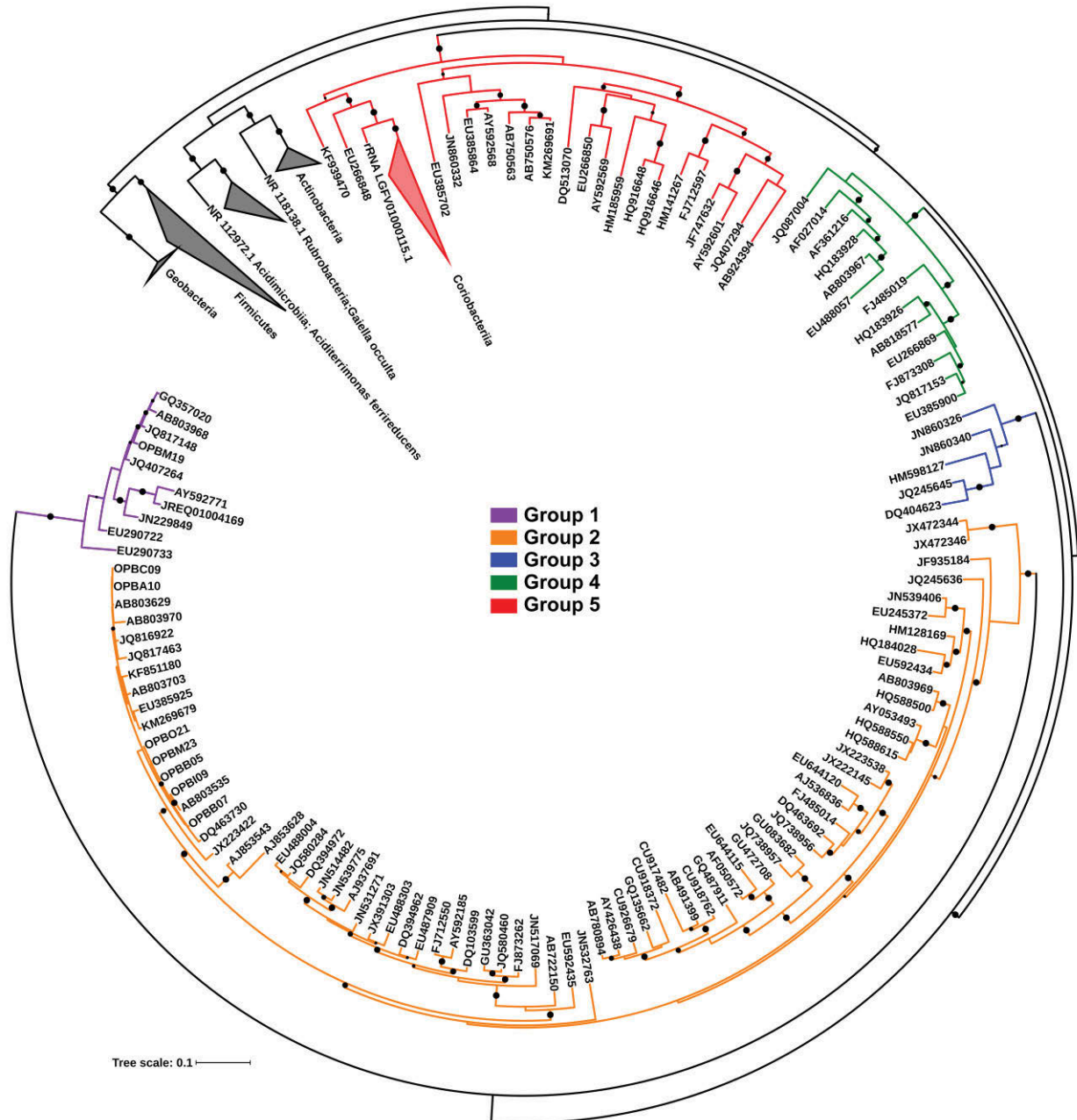


Figure 4.1: Maximum-likelihood tree based on representative members of actinobacterial classes and available full-length 16S rRNA sequence data from the “OPB41 clade” in the SSU Ref NR 99 database (labelled with colored arcs). Group designed were assigned to cluster sequences within the “OPB41 clade.” This tree was rooted with Geobacteria and Firmicutes sequences as an outgroup. The size of the black dots on tree branches is proportional to a Shimodaira-Hasegawa local support score from ranging from 0.8 to 1.

members of the OPB41 Group 2 (Bird et al, 2017, *in prep*). No full length 16S rRNA genes were available from SAG OPB41_O22; however, a 16S rRNA gene fragment was sequenced after primer amplification of the SAG DNA, and was a 100% match to the OPB41_M19 gene, placing it in Group 1. Percent identity between OPB41 Group 2 genomes was never less than 98.59% with 100% identity between full length sequences from OPB41_B05, OPB41_M23, OPB41_O21. OPB41_A10 and OPB41_C09 16S rRNA sequences were also 100% identical.

When employing the commonly used cutoff of 97% sequence identity for species level similarities each of the eight Group 1 SAGs would be considered the same species. In contrast, the percent identity between Group 1 and Group 2 sits at 83.17% and likely do not belong to the same genus. Below we discuss the mean average nucleotide identity methods which offer further evidence and detail into the delineation of species among the SAGs. Another 16S sequence from the assembled metagenomic contigs sampled at 31 mbsf in Canterbury Basin off the coast of New Zealand also belong to the OPB41 Group 1 (54). A 16S rRNA gene that binned with a MAG from a Canadian aquifer clusters with members of the cultured class, Coriobacteriia (12). The 16S rRNA gene from the Canadian aquifer as well as other members of OPB41 Group 5 are monophyletic with the previously described class, Coriobacteriia. Because limited genomic information is available for OPB41 Group 3-5, this study will focus primarily on OPB41 Group 1 and 2.

OPB41 Group 2 is a diverse group with the majority of the 16S rRNA genes found in anaerobic marine sediments. Further evidence that OPB41 is within the Actinobacteria comes from the presence of a conserved insertion sequence, which is a synapomorphic marker of Actinobacteria (Figure 4.2) (13). Group 1, 2, and 3 appear to be a separate class from the Coriobacteriia, both by their 16s rRNA gene placement and the fact that the OPB41 SAGs sequenced in this study and MAGs from Anantharaman et al (2016) lack conserved signature indels found in all Coriobacteriia such as the six amino acid deletion in adenylate kinase (Figure 4.3). These data suggest that these OPB41 genomes constitute a distinct uncultured class within the Actinobacteria.

Commonalities across environments where OPB41 16S rRNA genes are found hold clues to their potential metabolisms and limitations on their dispersal. Some members of OPB41 Group 1 are from a diverse group of anoxic and often deep marine sediment samples from the

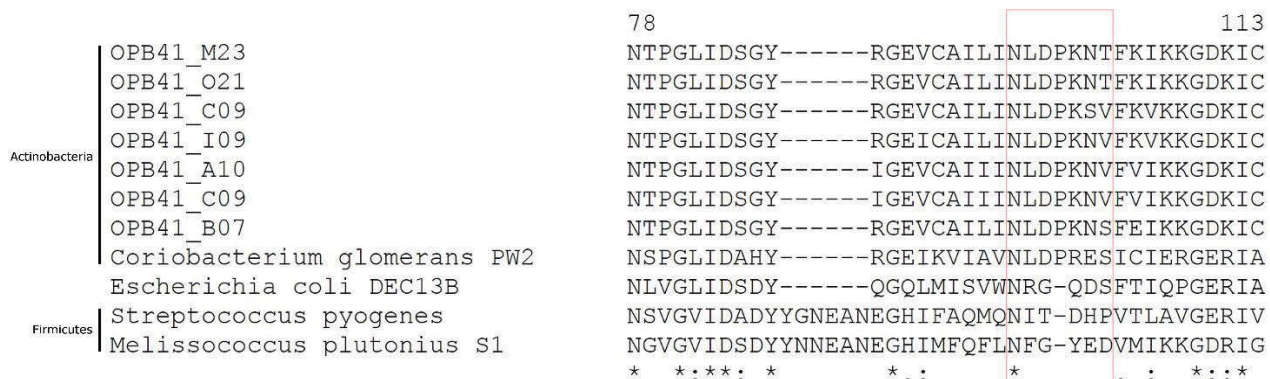


Figure 4.2: OPB41 SAGs contains a conserved signature indel present in all Actinobacteria deoxyuridine 5'-triphosphate (dUTP) nucleotidohydrolase protein sequences. The red box highlights the highly conserved containing the single amino acid indel common to all known Actinobacteria as characterized by Gupta et al (5).

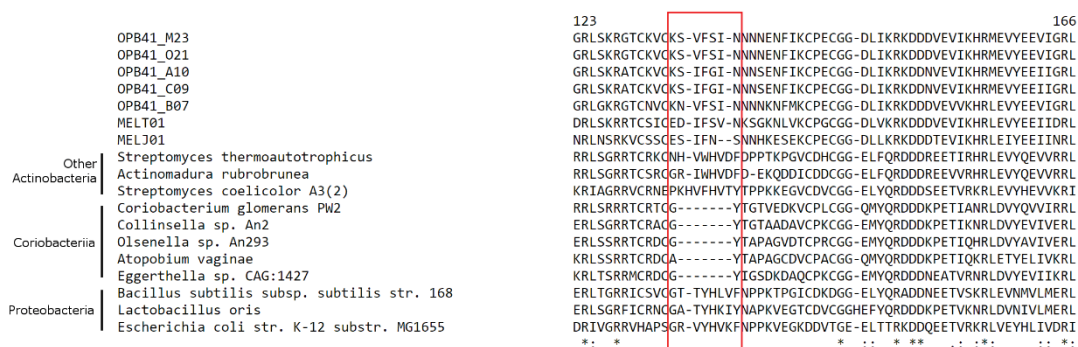


Figure 4.3: OPB41 SAGs do not contain a conserved signature indel present in all Coriobacteriia six amino acid deletion in adenylate kinase protein sequences. The red box highlights the amino acid indel common to all known Coriobacteriia as characterized by Gupta et al (2013).

costal margins of New Zealand, Ireland, Italy, Japan, and Namibia. Other Group 1 members were sampled from terrestrial mud volcanos where the source fluid was anoxic and contained abundant hydrocarbons. Within OPB41 Group 2, some subclades cluster based on environment type. Tight clusters formed in sequences sourced from anaerobic bioreactors, lake sediments, and deep anoxic ocean sediments. 16S rRNA genes from the SAGs in this study were found in one such tight cluster of deep anoxic marine samples including those near Shimokita Peninsula, Formosa Ridge and Good Weather Ridge in the South China Sea, and marine sediment samples near Hong Kong dated to the Pleistocene (NCBI direct submissions). The only non-marine member of this subclade was from the anoxic bottom waters of Lake Tanganyika in the Great Rift Valley of Africa (NCBI direct submission). Other frequent sources for sequences in Group 2 included hypersaline microbial mats, subsurface environments influenced by oil and gas, permafrost, and anoxic lake bottom waters (14–17). Our Baltic Sea deep subsurface 16S rRNA genes from Group 2 recruited more transcripts than Group 1 (Bird et al, 2017, *in prep*), suggesting that Group 2 may have been more abundant or active in situ. If this is true in other environments, this may explain why Group 2 sequences are more frequently recovered than Group 1 sequences.

OPB41 Group 3 contains 16S rRNA genes from diverse anoxic environments including contaminated sediments, mud volcanos, marine surface sediments, and the inner surfaces low temperature serpentinite chimneys (18–20). OPB41 Group 4 includes the original OPB41 16S sequence and sequences from other hot springs, freshwater and marine sediment, contaminant terrestrial aquifers, and a deep sinkhole associated with chemoautotrophic microbial mats (21–25). OPB41 Group 5 includes still more samples from marine sediment, environments influenced by oil and gas, deep aquifers, and hypersaline microbial mats (15). The most consistent environmental attribute across all five groups remains anoxia and being in the subsurface of Earth. These findings suggest OPB41 as a taxon is largely composed of strict anaerobes that may be outcompeted by other clades in surface-associated environments.

Given that OBP41 sequences are associated with anoxic environments, often found in mud volcanos and hot springs, they may have similar microbial dispersal mechanisms as other subsurface organisms. Hyperthermophilic firmicutes spores are lifted from the ocean floor and

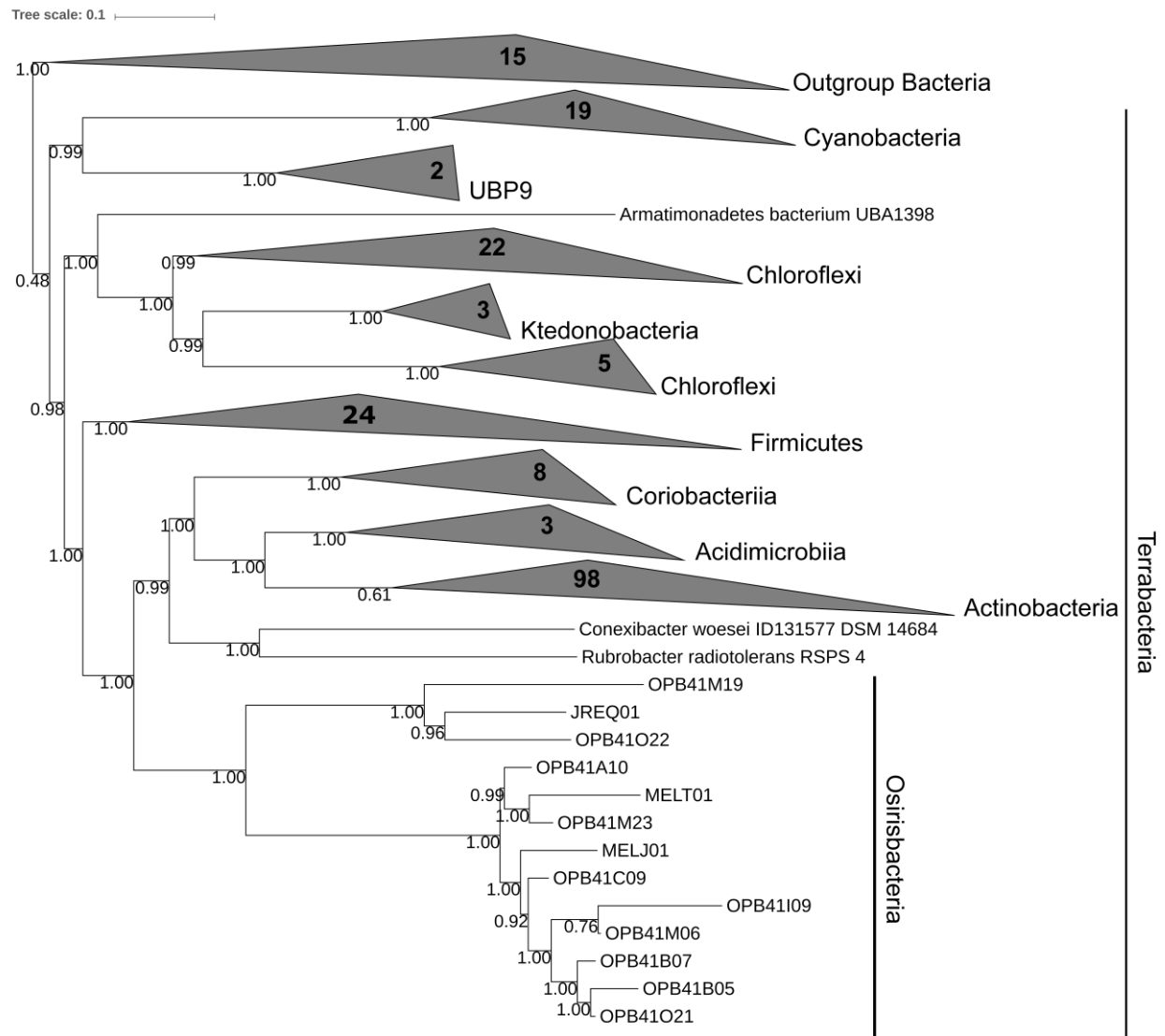


Figure 4.4: The maximum likelihood tree was inferred from the concatenation of 128 single-copy proteins and spans a set of representative genomes. Colored bins denote membership in the labeled bacterial classes within the Terrabacteria superphyla. The tree is rooted at the node leading to a diverse group of bacteria outside the Terrabacteria superphyla (purple). The numbers at each node of are the Shimodaira-Hasegawa local support scores.

carried on ocean currents by the outflow of hydrothermal vents (26, 27). Researchers also used catalyzed reporter deposition-fluorescence *in situ* hybridization to track the dispersal of Atribacteria populations from mud volcano plumes (28). Similarly, OPB41 taxa members may be dispersed as “deep-biosphere seeds”.

No other genomic information besides 16S rRNA genes is currently available for OPB41 Groups 3-5, so their exact placement within the Actinobacteria awaits further information. Considering only groups 1 and 2, the maximum likelihood tree inferred from single-copy proteins agrees with 16S rRNA gene data that OPB41 is a novel bacterial class within the phylum Actinobacteria (Figure 4.4). The tree showed deep-branching of the available OPB41 genomes away from other known Actinobacterial classes. The average branch length leading from Actinobacteria OPB41 to its closest relative *Connexibacteria woesii* of 0.78 substitutions per position (s/p) was within the range of genomic assemblies classified as new phyla in a recent study (7), and at least as long as any of those previously described as separate classes (Figure 4.4). OPB41 genomes branched even more deeply than recently described novel classes including: Rubrobacteria and Thermoleophilia, as well as, Coriobacteriia, which was previously described as the earliest branching class of Actinobacteria (29). However, given the presence of the synapomorphic indels discussed above, we do not propose that OPB41 is a new phylum. We instead propose that OPB41 be reclassified as a novel bacterial class with the name “Candidatus Osirisbacteria”. Osiris is the Egyptian god of the underworld, an apt namesake for this class of Actinobacteria which has largely restricted to anoxic subsurface environments.

Comparative Genomics

OPB41 genomes in this study have very low G+C contents of 30% for Group 1 and 35% for Group 2 (Figure 4.5). The majority of cultured Actinobacteria have high G + C content, with some greater than 70% (30). A review of Actinobacterial genomes stated that, with the exception of *Trophreyima whipplei*, all previously cultured Actinobacteria have G+C contents greater than 51% (30). However, researchers using “Actinobacteria-associated” metagenomic fragments predicted that low G + C content Actinobacteria are prevalent in freshwater environments (31). Their prediction of 37-43% G + C content for acI-B genomes prove true when researchers analyzed this freshwater taxon (32). Another study which produced 42 genomic assemblies from

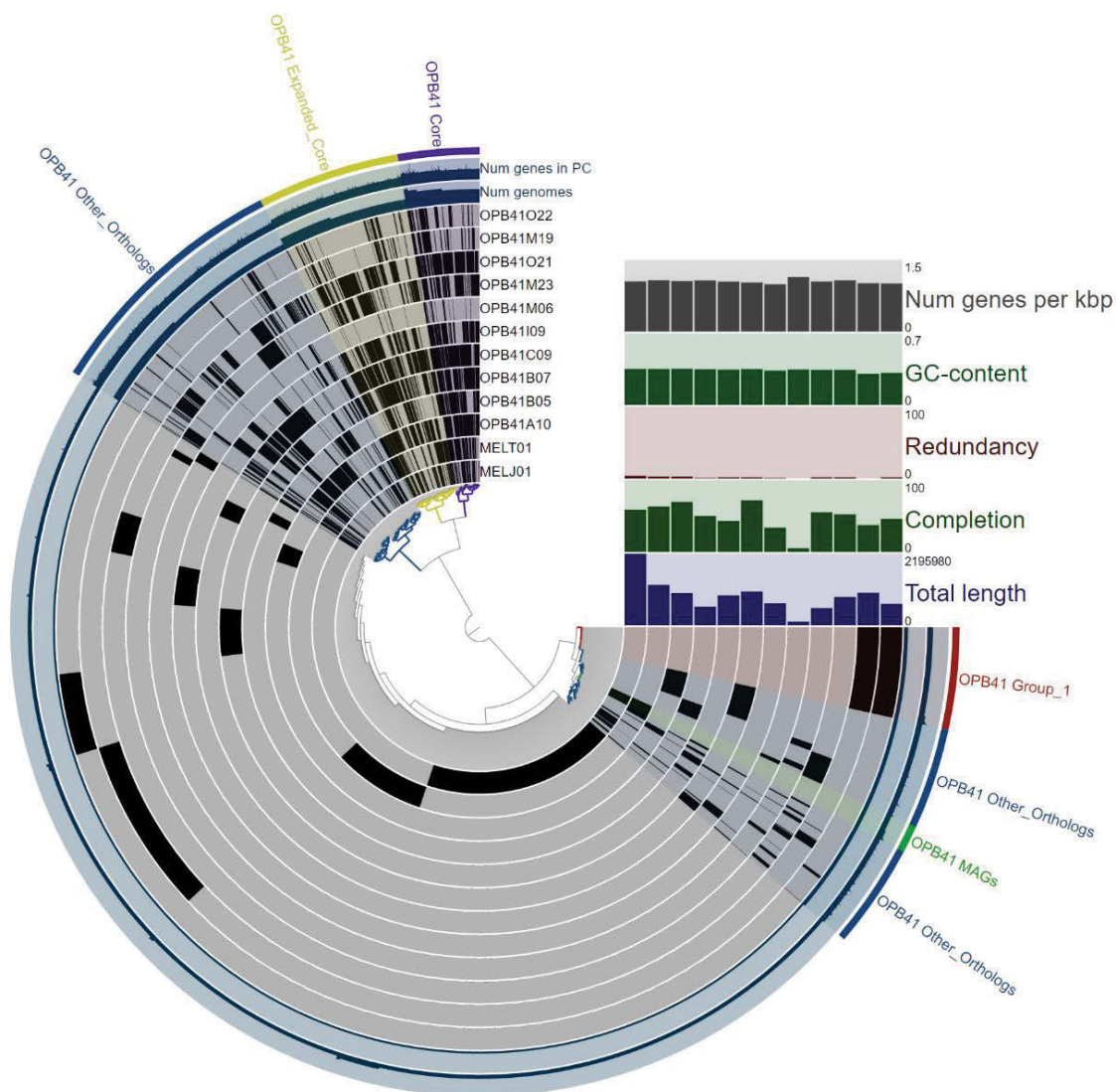


Figure 4.5: Comparative analysis of protein clusters within all known Actinobacteria genomes. The inner tree was constructed from matrix of presence:absence of aligned protein clusters across the genomes in this study. Black bars at the tips of inner tree represent the presence of absence of a protein within a genome. Colored bins along the edge highlight co-occurrence patterns across the genomes: OPB41_Group1 (red), OPB41_Core (purple), OPB41_Expanded_Core (light green), Other_Orthologs (blue), OPB41_MAG from the Rifle, Colorado (green), and Non_Orthologs (no label). Genome statistics for each genome are displayed in the colored barplot.

new actinobacterial classes reported a range of G + C content between 38 and 71% with a mean G + C of 50% (7). Even though these data show that environmental Actinobacteria can have lower G content than Actinobacterial cultures, Osirisbacteria represent a new endmember for low G + C content within the phyla. Wu et al linked the acquisition of DnaE2 within the Terrabacteria ancestor to increased G + C content, the acquisition of neofunctions, and the colonization of land by microbes (33). The lack of DnaE2 in Osirisbacteria could explain the surprisingly low G + C content and potentially its limited dispersal.

While 16S sequence similarity suggests species level similarities within Group 1 and Group 2, comparison of the nucleotide composition between the SAGs suggests further species differentiation. Microbial genomes are assumed to belong to the same species if they share an average nucleotide identity (ANI) of greater than 92.2% (34). In contrast to the analysis based on 16S percent identity which suggests all Group 2 SAGs belong to the same species, the dRep genome comparison tool suggested OPB41_B05, OPB41_M23, OPB41_O21, and OPB41_B07 SAGs belong to one species, with OPB41_B07 being from a different strain (ANI < 99%), and OPB41_A10 and OPB41_C09 (ANI < 92.2%)(34, 36). OPB41_I09 and OPB41_M06 did not contain sufficient quantities of overlapping genomic information (MASH Average Nucleotide Identity < 90%) for the gANI algorithm to make an accurate prediction of species relatedness. While the available portions of the 16S rRNA genes of OPB41_O22 and OPB41_M19 were identical, their ANI value suggested they do not belong to the same species. Taken together, the SAGs can be assumed to be from organisms of two species within Group 1 and at least two species within Group 2.

Analysis of the overlapping genomic regions between Group 1 and Group 2 OPB41 genomes yielded a surprisingly small list of shared genomic features, given their high similarity of evolutionarily conserved genes. Overall, only 13 genes were shared between the two Group 1 genomes and the eight Group 2 genomes in this study. In total 1153 orthologs were discovered in Group 1, while Group 2 contained 255 such orthologs. While zero genes were shared across all 8 aligned regions of the fragmented Group 2 genomes, the number of orthologs increased with decreased match number (Table 4.1). There were considerably fewer total basepairs from the OPB41_M06 genome than from each of the other Group 2 genomes; however, even after decreasing the overlap requirement to 7 genomes, only 28 orthologs were observed. Overall,

Table 4.1: Level of overlap between orthologs in Actinobacteria OPB41

	Orthologs in Group 2	Orthologs in Group 1	Overlap
non-orthologs	2147	1054	3201
At least 2	1153	255	13
At least 8	0		
At least 7	28		
At least 6	85		
At least 5	195		
At least 4	355		
At least 3	588		

69.5% of coding sequence (CDS) features contained no orthologs within the genomes in this study.

The 13 orthologs found between the Group 1 and Group 2 fell on two overlapping sequence regions. One overlapping region containing five orthologs resembled a ribose sugar transport system with an annotated D-ribose binding periplasmic protein RbsB, 2 Ribose transport system permease protein RbsC, an ATP-binding RbsA, and another ATP-binding protein similar to the MalK. *E. coli* with mutant RbsB proteins do not have the ribose mediated taxis phenotype (36). Previous analyses of enzymatic activities and transcript recruitment to these genomes suggest the Actinobacteria OPB41 in these sediments participate in the degradation of D-linked sugars (10). Overall, these findings suggest that Actinobacteria OPB41 contains a conserved ribose sugar transport system that degrades D-linked sugars.

Another overlapping region contains orthologs annotated as a hypothetical protein, a DAK2 domain, an ATP-dependent DNA helicase RecG, a ribosomal RNA small subunit methyltransferase D, a region which is inverted in Group 1 containing a hypothetical protein and Phosphopantetheine adenylyltransferase coaD, 50S ribosomal protein L32, succinyl-CoA synthetase α subunit, and another hypothetical protein. Two of the genes in the overlapping region are homologous subunits of protein complexes important in the tricarboxylic acid (TCA) cycle and a third is homologs to a component of the large subunit of the ribosome. Overall, it is unclear what physiological role these conserved genes may play as together they do not resemble a previously described protein complex.

The 28 orthologs found in seven overlapping regions of OPB41 Group 2 members include four sets of adjacent proteins with functions including amino acid and vitamin synthesis, protein translation, stress response and energy conservation. Genes annotated as 30S ribosomal protein S2, elongation factor Ts, uridylate kinase, and ribosome-recycling factor were in the first set of genes. These genes are likely involved in protein translation. Another set of genes included a mix of genes for biosynthesis of either amino acids or vitamins and included a homolog to a stress repressor regulatory gene. Argininosuccinate synthase and argininosuccinate lyase which carry out part of the pathway to synthesize L-arginine were counted among the orthologs. Finally, homologous proteins similar to the four subunit Sulphydrogenase complex described in

Pyrococcus furiosus. In *P. furiosus*, this complex is capable of accepting electrons from NADPH and reducing protons to H₂ gas playing a key role in the energetics of fermentation (37).

With analysis of overlapping genomic regions yielding few similarities between predicted metabolisms, comparative analysis of the predicted proteome allowed for the description of core proteins conserved across the OPB41 genomes. Two other genomic assemblies from metagenomes sourced from uranium contaminated aquifers having similar conserved proteins in our conserved indel analysis were also included in this pangenomic analysis (Figure 4.5). The number of protein clusters for comparison increased to 1658 across seven genomes (OPB41_Core) from 1153 orthologs across any two genomes in Group 2 (Table 4.1). Here we discuss key metabolisms inferred sets of overlapping protein clusters.

OPB41_Core proteins were enriched in COG categories C, J, I, L, O, and U relative to the Non_Orthologs and Other_Orthologs sets (Figure 4.6). Category C concerns energy conversation and production. This includes COGs matching a bifunctional CO dehydrogenase/acetyl-CoA synthase complex, which could carry out the carbon monoxide side of the Wood-Ljungdahl pathway (25). Also among these orthologs were subunits for succinate dehydrogenase/fumarate reductase which is important for anaerobic growth in *E. coli* and other pieces of the electron transport chain. The Expanded_Core included more COGs involved in the electron transport chain including heterodisulfide reductase, components of hydrogenases including a small sub-unit of Ni,Fe-hydrogenase, ATPase subunits, and COG matching electron acceptors such as Ferredoxin, Rubrerythrin, and Rubredoxin. Alcohol dehydrogenase, a key enzyme in alcohol fermentation were also found in Expanded_Core. Although succinate dehydrogenase/fumarate reductase found in the OPB41_Core, other protein complexes within OPB41_Core did not suggest a common potential terminal electron donor (e.g. sulfate). Absent are cytochromes or other components of an aerobic respiration.

Protein clusters of OPB41_MAG appear to be enriched in carbohydrate transport and metabolism COGs. The apparent enrichment in the carbohydrate transport and metabolism COGs is largely driven by sets of orthologs encoding ABC-type glycerol-3-phosphate transport systems (Figure 4.6). These transport systems are often flanked by extracellular hydrolases such as alpha and beta-galactoside, beta-glucosidase, hexokinase, and other diverse hydrolases targeting both disaccharides and ribose sugars. These transport systems and hydrolases are also

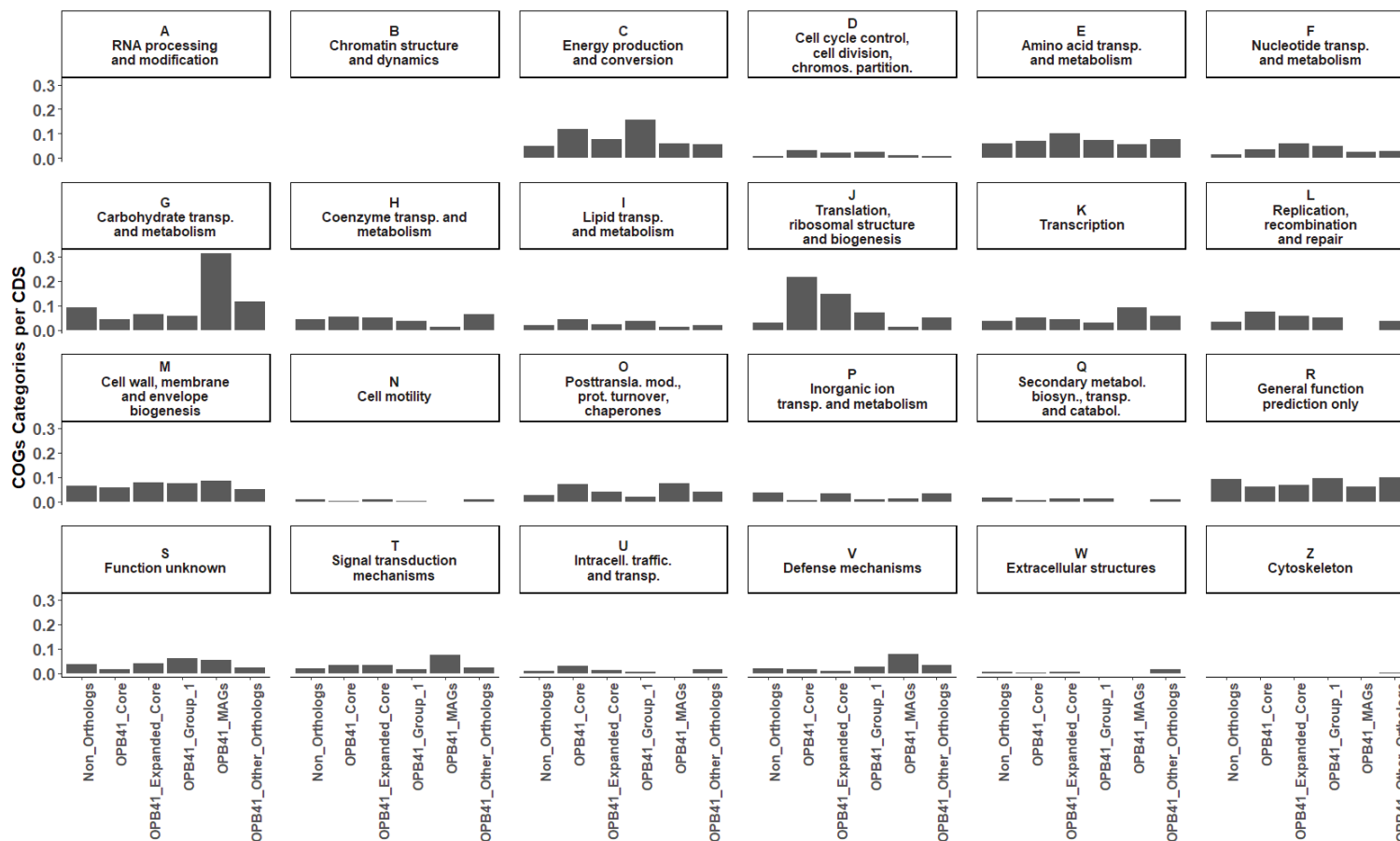


Figure 4.6: The proportion of COG categories varies across the sets of protein clusters on the x-axes. The bar graphs denote the number of coding sequences (CDS) in each COG categories relative to the total number of CDS in each set of protein clusters. COG categories correspond to proteins involved in different classes of cellular metabolism.

encoded by OPB41 Group 1 and Group 2; however, a greater number and diversity of hydrolases were observed in the MAGs and none of the transport systems were found in the OPB41_Core or OPB41_Expanded_Core. It is possible that the OPB41 members at the Rifle, Colorado site import and catabolize a greater diversity of organic sugars. It has been observed that organic carbon pools tend to become less abundant and more refractory as depth in sediment increases (38). The total organic carbon in the Baltic sediment site decreased with depth (39).

Additionally, metabolite profiles contained only one detected disaccharide, trehalose (10). COGs for periplasmic beta-glucosidase and related glucosidases as well as cellobiose phosphorylase, which functions as a cellulase in other bacteria, were present in the OPB41_Expanded_Core. Previous analysis of transcriptomes from the Expedition 347 showed transcript recruitment that correlated with measured specific enzyme activities for beta-glucosidase and cellulase (10).

OPB41_MAG protein clusters also appear to be enriched in defense mechanisms COGs. Two such defense mechanism COGs, RidA family proteins and lipoproteins NlpE involved in copper resistance, were found only in the OPB41_MAG protein cluster set. Additionally, while multidrug efflux pump subunits were found in the SAGs, several orthologous pairs were unique to the terrestrially sourced MAGs. Greater competition of among OPB41 members at the Rifle, Colorado site may explain the increased diversity of defense mechanisms.

Actinobacteria as a phyla has been a rich source of value-added products and are known for an expansive array of secondary metabolisms (40). While secondary metabolite production seems to be unlikely in the energy-limited environments that are rich in Actinobacteria OPB41, surface marine sediments have been a frequent source of antimicrobial product producing Actinobacteria cultures (41, 42). Secondary metabolite production involves the expression of protein clusters which transform biomolecules into secondary metabolites that may have antibiotic or other properties that are useful to humans and industry. The online tool antiSMASH is a bioinformatic tool that can identify secondary metabolite clusters within bacterial, yeast, and plant genomes. Each of the annotated Actinobacteria OPB41 SAGs in this study were submitted for analysis at <http://antismash.secondarymetabolites.org>. Only one predicted gene cluster was found among the OPB41 SAGs containing a single 'core biosynthetic gene', a Long-chain-fatty-acid--CoA ligase. In MELJ01 three predicted Ladderane synthesis clusters were found. Each

MELJ01000252

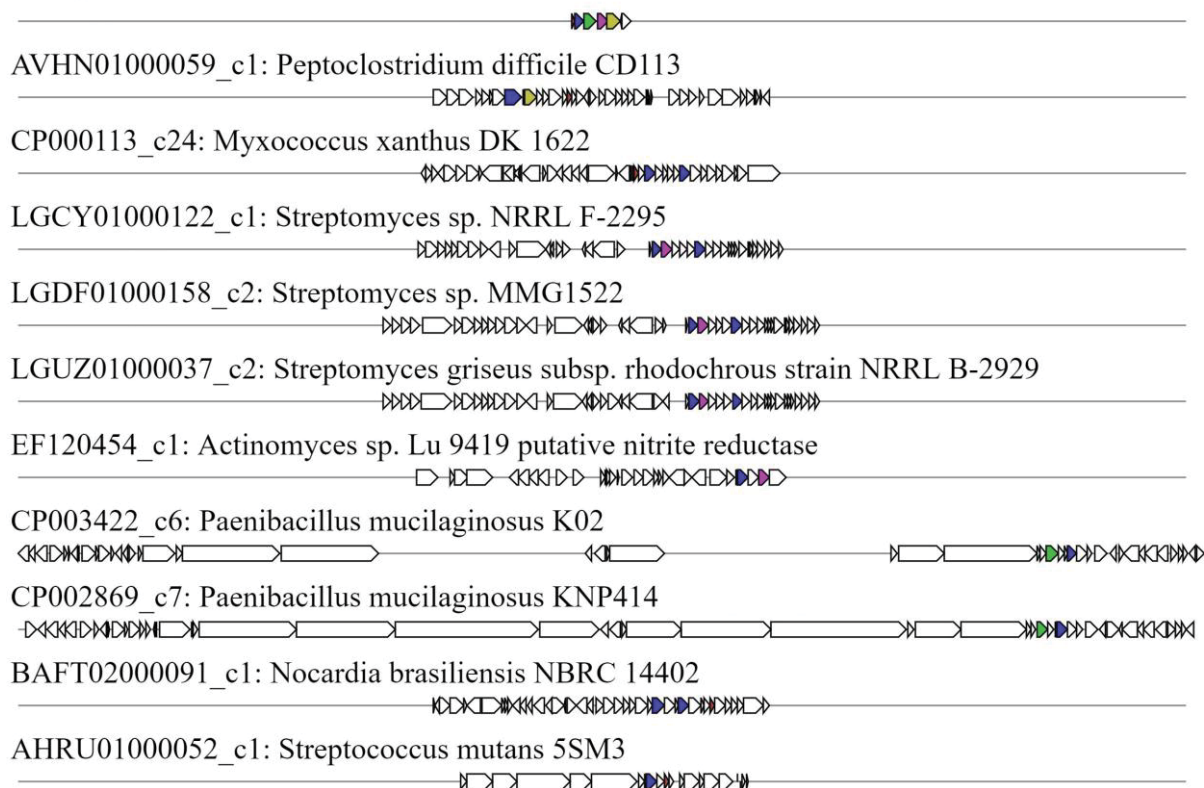


Figure 4.7: Alignment of the predicted ladderane biosynthetic gene clusters. Matching colored triangles represent gene orthologs to include: UDP-3-O-acetyl-N-acetylglucosamine deacetylase (orange), Beta-ketoacyl synthases with PKS_KS active site (blue), dehydrogenase (green), Beta-ketoacyl synthases (pink), and another dehydrogenase (gold). This alignment is one of three such biosynthetic gene clusters identified MELJ01 MAG using antiSMASH.

contained one to two beta-ketoacyl synthases flanked by UDP-3-O-acyl-N-acetylglucosamine deacetylase and dehydratase genes (Figure 4.7). The absence of abundant biosynthetic secondary metabolite gene clusters stands in contrast to the large array of multidrug efflux pumps that are encoded in these genomes. However, with a lack of polyketide synthases in the SAGs or other known antibiotic production genes it is unclear what purpose these efflux pumps serve other than defense against other diverse community members or transport of some unknown metabolite. Subsurface marine sediment environment such as the Baltic sea sediment are characterized as low energy environment with cell abundances that decrease with depth; therefore, it is likely that competition for physical space is not a key pressure on the community (43, 44).

Materials & Methods

Sample collection

Deep subsurface sediment cores were collected during IODP Expedition 347: Baltic Sea Paleoenvironment with samples for single cell sorting taken from M0060 at 37 and 85 meters below the seafloor (mbsf) and from M0059 at 41 and 65 mbsf. Detailed cell extraction methods were previously detailed in Bird et al, 2017, *in prep* (10). Sediments samples were assessed for drill fluid contamination and results were published along with the initial report (39).

Genome Sequencing and Assembly

Two separate sequencing runs were used to produce the Illumina paired-end reads for the SAGs in the study. As reported previously 2x125 basepairs (bp) reads were generated with the HiSeq platform at Molecular Research DNA Laboratory, Shallowater, TX using the Nextera read library preparation for SAGs, OPBI09 and OPBM19. The remaining genomic reads were 2x250 bp reads generated with the MiSeq platform at the University of Tennessee, Knoxville, TN using the same read library preparation. The resulting reads were trimmed with the Trimmomatic software with the following parameters: ILLUMINACLIP:Nextera-PE.fa:2:30:12:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:10:20 MINLEN:75 (45). The resulting trimmed reads were assembled with SPAdes 3.6 assembler with the following parameters: `spades.py -m 62 -o <OUTPUT_DIR> -sc -12 <PAIRED_READS> -s <SINGLES> -t 15 -k 21,33,55,77,99,127 -careful` (46). Those contigs matching the PhiX internal standard and those less than 1000 bp or having less than 5x read coverage were removed from the final assemblies. Genomes were annotated using the methods previously described (10).

The genome alignment tool MAUVE was used to analyze the overlapping regions between Group 1 and Group 2 OPB41 using the default progressive alignment settings. Sets of orthologous genes were defined within these aligned regions. In addition, coding sequences (CDS) were annotated using the cluster of orthologous genes (COG) database using the diamond alignment algorithm within the Anvi'o (47, 48). The Anvi'o pangenomics workflow (<http://merenlab.org/2016/11/08/pangenomics-v2/>) was used to define sets of proteins clusters. The cutoff for inclusion in the OPB41_Core and OPB41_Expanded_Core was the presents of an orthologous gene in 7 or 5, respectively, of the 12 OPB41 genomes. Proteins clusters found in 2 to 4 genomes were included in the OPB41_Other_Orthologs set. Other protein clusters included orthologous genes unique to Group 1 genomes or the metagenomic genomes from Anantharaman et al (11). Protein clusters without matches were defined as Non_Orthologs. Secondary metabolite syntheses cluster were identified using the AntiSmash web server (49).

Phylogenomics

Full-length 16S rRNA genes derived from SAGs were identified by barrnap (50). 16S sequences that cluster in the "OBP41 clade" in SILVA SSU ref 99 were exported from the ARB platform. (51, 52). These exported sequences were aligned with sequences from the OPB41 SAGs and representative sequences of Firmicutes and Geobacter drawn from NCBI using MAFFT with the -auto setting. (53) Some of these sequences were directly submitted to NCBI and have not yet been published formally. The tree was rooted with Archaeopteryx 0.9920 at the branch leading to Geobacter, (54) and was then visualized and edited for publication using the Anvi'o.

Single copy conserved proteins OBP41, other members of Terrabacteria superphylum, and a few distantly related bacterial phyla sourced from NCBI Assembly and WGS databases. In order to included additional genomic information from (JREQ000000000.1) Actinobacteria within the Canterbury Basin marine metagenomes, the VizBin software was used to quickly bin a set of genomic information including single-copy genes related to OPB41 (Figure S4.1) (55, 56). Single-copy conserved genes were identified using the hmmer 3.0 alignment algorithm within Anvi'o (57). The 126 proteins first identified as conserved by Rinke et al were aligned and concatenated (2). Aligned and concatenated single-copy genes were trimmed using trimAI using the strictplus parameter removing 77,032 amino acid positions and leaving 19,518 positions for tree construction (58). Phylogenetic tree construction was performed with

FastTree using the following parameters: -cat 20 -gamma (59). The final tree (Fig. 4) was modified using iTOL (60).

The genome comparison tool dRep was utilized in order to level the fast and accurate MASH and gANI algorithms for generating average nucleotide identity comparisons across the genomes (34,35). The MASH algorithm is first used to sort genomes into sufficiently similar cluster before performing accurate average nucleotide identity assessments of species similarity using the gANI algorithm. The resulting ANI values were then used to infer the probable taxonomic classification shared by the SAGs in the study.

Each of the figures in this were edited for publication using the GIMP 2.0 software. The graphical R package ggplot2 was also used for data visualization.

Acknowledgments

Shipboard Science Party of IODP Expedition 347. Justin and Lori Bird for writing assistance, Dan Williams and members of the Terry Hazen and Frank Löffler labs for help with library prep and sequencing. Funding: NSF-OCE, C-DEBI biomolecular grant, Simons (for Jordan's salary).

List of References

1. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR. 1998. Novel Division Level Bacterial Diversity in a Yellowstone Hot Spring. *J Bacteriol* 180:366–376.
2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499:431–437.
3. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523:208–211.
4. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hermsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nat Microbiol* 1:doi: 10.1038/nmicrobiol.2016.48.
5. Baker BJ, Lazar CS, Teske AP, Dick GJ. 2015. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* 3:doi: 10.1186/s40168-015-0077-6.
6. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541:353–358.
7. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2:1533–1542.
8. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 7:13219.

9. Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol* 4:44.
10. Bird JT, Tague ED, Zinke LA, Schmidt JM, Steen AD, Reese BK, Marshall IPG, Webster G, Weightman A, Castro HF, Campagna SR, Lloyd KG. 2017. Atribacteria support community-wide microbial persistence through 44,000 years of Baltic Sea sedimentation. Prep.
11. Buongiorno J, Turner S, Webster G, Asai M, Shumaker AK, Roy T, Weightman A, Schippers A, Lloyd KG. 2017. Interlaboratory quantification of Bacteria and Archaea in deeply buried sediments of the Baltic Sea (IODP Expedition 347). *FEMS Microbiol Ecol* 93.
12. Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, Andersen GL, Banfield JF. 2016. Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and Other Microbial Community Members in Biogeochemical Transformations in Oil Reservoirs. *mBio* 7:e01669-01615.
13. Gupta RS, Chen WJ, Adeolu M, Chai Y. 2013. Molecular signatures for the class Coriobacteriia and its different clades; proposal for division of the class Coriobacteriia into the emended order Coriobacteriales, containing the emended family Coriobacteriaceae and Atopobiaceae fam. nov., and Eggerthellales ord. nov., containing the family Eggerthellaceae fam. nov. *Int J Syst Evol Microbiol* 63:3379–3397.
14. Liebner S, Harder J, Wagner D. 2008. Bacterial diversity and community structure in polygonal tundra soils from Samoylov Island, Lena Delta, Siberia. *Int Microbiol Off J Span Soc Microbiol* 11:195–202.
15. Kirk Harris J, Gregory Caporaso J, Walker JJ, Spear JR, Gold NJ, Robertson CE, Hugenholtz P, Goodrich J, McDonald D, Knights D, Marshall P, Tufo H, Knight R, Pace NR. 2013. Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME J* 7:50–60.
16. Acosta-González A, Rosselló-Móra R, Marqués S. 2013. Characterization of the anaerobic microbial community in oil-polluted subtidal sediments: aromatic biodegradation potential after the Prestige oil spill. *Environ Microbiol* 15:77–92.

17. Lanoil BD, Sassen R, La Duc MT, Sweet ST, Nealson KH. 2001. Bacteria and Archaea physically associated with Gulf of Mexico gas hydrates. *Appl Environ Microbiol* 67:5143–5153.
18. Abulencia CB, Wyborski DL, Garcia JA, Podar M, Chen W, Chang SH, Chang HW, Watson D, Brodie EL, Hazen TC, Keller M. 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. *Appl Environ Microbiol* 72:3291–3301.
19. Cheng T-W, Chang Y-H, Tang S-L, Tseng C-H, Chiang P-W, Chang K-T, Sun C-H, Chen Y-G, Kuo H-C, Wang C-H, Chu P-H, Song S-R, Wang P-L, Lin L-H. 2012. Metabolic stratification driven by surface and subsurface interactions in a terrestrial mud volcano. *ISME J* 6:2280–2290.
20. Li J, Peng X, Zhou H, Li J, Sun Z. 2013. Molecular evidence for microorganisms participating in Fe, Mn, and S biogeochemical cycling in two low-temperature hydrothermal fields at the Southwest Indian Ridge. *J Geophys Res Biogeosciences* 118:665–679.
21. Sahl JW, Fairfield N, Harris JK, Wettergreen D, Stone WC, Spear JR. 2010. Novel microbial diversity retrieved by autonomous robotic exploration of the world's deepest vertical phreatic sinkhole. *Astrobiology* 10:201–213.
22. Marteinsson VT, Hauksdóttir S, Hobel CF, Kristmannsdóttir H, Hreggvidsson GO, Kristjánsson JK. 2001. Phylogenetic diversity analysis of subterranean hot springs in Iceland. *Appl Environ Microbiol* 67:4242–4248.
23. Tischer K, Kleinstaub S, Schleinitz KM, Fetzner I, Spott O, Stange F, Lohse U, Franz J, Neumann F, Gerling S, Schmidt C, Hasselwander E, Harms H, Wendeberg A. 2013. Microbial communities along biogeochemical gradients in a hydrocarbon-contaminated aquifer. *Environ Microbiol* 15:2603–2615.
24. Sessitsch A, Weilharter A, Gerzabek MH, Kirchmann H, Kandeler E. 2001. Microbial population structures in soil particle size fractions of a long-term fertilizer field experiment. *Appl Environ Microbiol* 67:4215–4224.

25. Liu J, Wu W, Chen C, Sun F, Chen Y. 2011. Prokaryotic diversity, composition structure, and phylogenetic analysis of microbial communities in leachate sediment ecosystems. *Appl Microbiol Biotechnol* 91:1659–1675.
26. Hubert C, Loy A, Nickel M, Arnosti C, Baranyi C, Brüchert V, Ferdelman T, Finster K, Christensen FM, Rezende JR de, Vandieken V, Jørgensen BB. 2009. A Constant Flux of Diverse Thermophilic Bacteria into the Cold Arctic Seabed. *Science* 325:1541–1544.
27. Müller AL, de Rezende JR, Hubert CRJ, Kjeldsen KU, Lagkouravdos I, Berry D, Jørgensen BB, Loy A. 2014. Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *ISME J* 8:1153–1165.
28. Hoshino T, Toki T, Ijiri A, Morono Y, Machiyama H, Ashi J, Okamura K, Inagaki F. 2017. Atribacteria from the Subseafloor Sedimentary Biosphere Disperse to the Hydrosphere through Submarine Mud Volcanoes. *Front Microbiol* 8: Article 1135
29. Verma M, Lal D, Kaur J, Saxena A, Kaur J, Anand S, Lal R. 2013. Phylogenetic analyses of phylum Actinobacteria based on whole genome sequences. *Res Microbiol* 164:718–728.
30. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, Sinderen D van. 2007. Genomics of Actinobacteria: Tracing the Evolutionary History of an Ancient Phylum. *Microbiol Mol Biol Rev* 71:495–548.
31. Ghai R, McMahon KD, Rodriguez-Valera F. 2012. Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environ Microbiol Rep* 4:29–35.
32. Ghylis TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, Sczyrba A, Stepanauskas R, Grossart H-P, Woyke T, Warnecke F, Malmstrom R, Bertilsson S, McMahon KD. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater actI Actinobacteria lineage. *ISME J* 8:2503–2516.
33. Wu H, Fang Y, Yu J, Zhang Z. 2014. The quest for a unified view of bacterial land colonization. *ISME J* 8:1358–1369.
34. Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*.

35. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. 2015. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 43:6761–6771.
36. Galloway DR, Furlong CE. 1977. The role of ribose-binding protein in transport and chemotaxis in *Escherichia coli* K12. *Arch Biochem Biophys* 184:496–504.
37. Silva PJ, van den Ban ECD, Wassink H, Haaker H, de Castro B, Robb FT, Hagen WR. 2000. Enzymes of hydrogen metabolism in *Pyrococcus furiosus*. *Eur J Biochem* 267:6541–6551.
38. Burdige DJ. 2005. Burial of terrestrial organic matter in marine sediments: A re-assessment. *Glob Biogeochem Cycles* 19:GB4011.
39. Andrén T, Jørgensen BB, Cotterill C, Green S, and the Expedition 347 Scientists. Expedition 347 of the mission-specific drilling platform from and to Kiel, Germany Sites M0059–M0067 12 September–1 November 2013. Integrated Ocean Drilling Program.
40. Solanki R, Khanna M, Lal R. 2008. Bioactive compounds from marine actinomycetes. *Indian J Microbiol* 48:410–431.
41. Claverías FP, Undabarrena A, González M, Seeger M, Cámara B. 2015. Culturable diversity and antimicrobial activity of Actinobacteria from marine sediments in Valparaíso bay, Chile. *Front Microbiol* 6: Article 737.
42. Dhakal D, Pokhrel AR, Shrestha B, Sohng JK. 2017. Marine Rare Actinobacteria: Isolation, Characterization, and Strategies for Harnessing Bioactive Compounds. *Front Microbiol* 8: Article 1106.
43. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A* 109:16213–16216.
44. Hoehler TM, Jørgensen BB. 2013. Microbial life under extreme energy limitation. *Nat Rev Microbiol* 11:83–94.
45. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,

- Alekseyev MA, Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 19:455–477.
47. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–60.
 48. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:doi: 10.7717/peerj.1319.
 49. Weber T, Blin K, Duddela S, Krug D, Kim HU, Brucoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43:W237–W243.
 50. Seemann T. 2013. barrnap 0.5: rapid ribosomal RNA prediction.
 51. Westram R, Bader K, Pruesse E, Kumar Y, Meier H, Glöckner FO, Ludwig W. 2011. ARB: a software environment for sequence data, p. 399–406. *In* *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* ed. de Bruijn FJ (Hoboken, NJ: John Wiley & Sons).
 52. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.
 53. Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
 54. Han MV, Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10:356.
 55. Gaboyer F, Burgaud G, Alain K. 2015. Physiological and evolutionary potential of microorganisms from the Canterbury Basin seafloor, a metagenomic approach. *FEMS Microbiol Ecol* 91.
 56. Laczny CC, Pinel N, Vlassis N, Wilmes P. 2014. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Sci Rep* 4:srep04516.
 57. Eddy SR. 2011. Accelerated Profile HMM Searches. *PLOS Comput Biol* 7:doi: 10.1371/journal.pcbi.1002195.

58. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
59. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5.

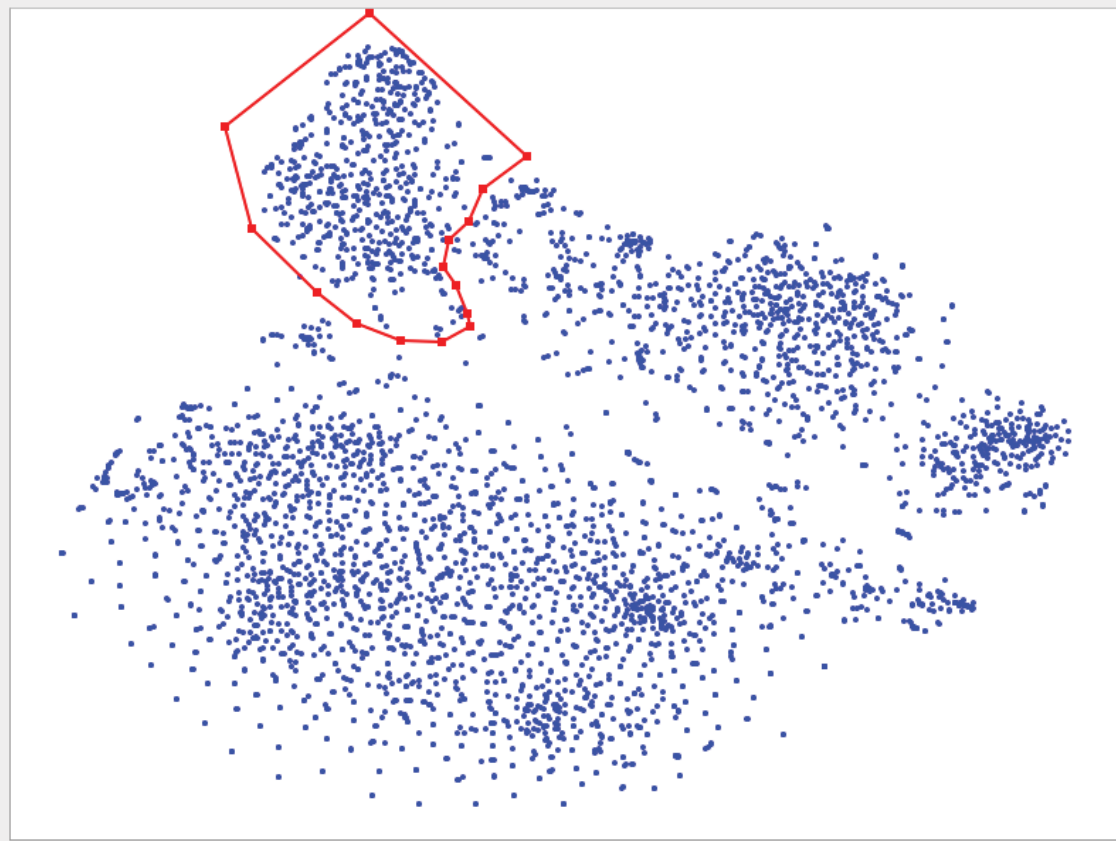


Figure S4.1: Two-dimensional scaling of Barnes-Hut Stochastic Neighbor Embedding (BH-SNE) of centered log-ratio (CLR)-transformed oligonucleotide signatures of assembled sequence fragments from metagenomic data from Canterbury Basin sediment, New Zealand (Accession: JREQ00000000.1) (55). The blue dots represent assembled contigs and the red polygon indicates the selection chosen which contained phylogenetic marker genes similar to other Actinobacteria OPB41.

Chapter V: Conclusion

Metagenomic binning and single-cell genomics have transformed our understanding of microbial phylogenetic and functional diversity in recent years. As we apply these methods to new environments or reanalyze previously acquired datasets, the number of new proposed phyla, classes, and orders continues to grow. Here I draw conclusions from the genomic analyses of a few of the 2.9×10^{29} cells within the marine sediment biome (1). Additionally, I offer some suggestions for future research which would build on these findings.

Current phylogenetic models have indicated that Altiarchaeales is one of the deepest branching groups in the Euryarchaeota phylum. Other researchers have suggested that they may even belong to a separate phyla (2, 3). While more genomes of Altiarchaeales and other deeply branching phyla radiations will be needed for a definitive phylogenetic assessment of Altiarchaeales, applying split gene and single-copy protein analyses to our current dataset has allowed us to tease out the relationship between the Alti-1 and Alti-2 groups.

Alti-2 is a diverse group, and it has been able to adapt to diverse environments including environments which lack the high flow rates associated with springs (Figure 2.4A). Ancestors of the Altiarchaeales that developed hami may have been able to persist in high-flow spring environments long enough for spring-specific adaptations to emerge. Surprisingly, the distribution of Alti-1 16S rRNA genes suggests the clade is now narrowly adapted to springs and maybe be unable to live in the environments where Alti-2 are found. Additionally, the Alti-2 genome associated with springs contained homologs to hamus structure proteins. Future research may focus questions on a potential single gene evolutionary sweep caused by the acquisition of hami. Additionally, future researchers may use the evolutionary relationship between Alti-1 and Alti-2 inferred in this study to investigate the effect of specialization for high vs. low retention rate environment on genetic diversity.

In chapters II and IV, synapomorphic mutations helped to establish a clear phylogenetic relationship between distantly related members of taxa which shared very few genes. This work adds to a growing body of research which employed similar methods to addresses similar problems (4, 5). With the use of conserved indels or apparent gene fission/fusion events, future researchers can find additional evidence to support relationships between other proposed members of a microbial taxa. These methods may be particularly useful with the recent expansion of the tree life because members of novel uncultured taxa often sit at the end of much

longer branches with fewer close relative than bacterial and archaeal taxa with cultured members.

The uncultured class Osirisbacteria, within the Actinobacteria, is potentially important to understanding the radiation and dispersal of the Terrabacteria superphylum. Patterns of dispersal inferred from 16S clone libraries suggest Osirisbacteria has some mechanism of dispersal which allow it to reach the narrow yet globally distributed anoxic environments it inhabits.

Osirisbacteria's association with mud volcanoes suggest it may have similar dispersal patterns to some types of Atribacteria (6). Additionally, Osirisbacteria lacks the error-prone polymerase, dnaE2, that previous researchers suggest was a key acquisition in life march from the sea to land (7). Future researchers could build on the findings presented here by asking question about what this earliest branching class of Actinobacteria can tell us about the dispersal of “deep-sea seeds” and the key functions for microbial colonization of Earth's landmasses.

I suggest that the uncultured phylum Atribacteria contains keystone species since they expresses the genes necessary to produce all 20 of the amino acids necessary for protein production and also contain amino acid export machinery that is promiscuous in its specificity of amino acids (8). With such a high proportion of the energy demand in persisting microorganisms going to protein repair, amino acids exported by Atribacteria may play a key role in the survival of other community members. It remains to be seen whether patterns of Atribacterial species and its co-occurring communities fit the macroecological theory which defines a keystone species (9, 10).

The predominance of protein domains associated with transcriptional and translational regulations (Figure 3.4) points to the potential role of gene regulation in microbial persistence. Some environments may select for persistence adaptations including the selective expression of certain genes. Future researchers may uncover specific gene regulatory networks with resulting persistence phenotypes that are key to survival in marine sediments.

The metabolites identified in the Baltic Sea sediment overlapped in their association with known microbial persistence mechanisms. Trehalose is a known osmoprotectant, important in preventing the denaturation of proteins, as well as a potential food source for organisms in the environment including Atribacteria (11, 12). Allantoin has been shown to cause global downregulation of transcription when administered as a drug and may provide Atribacteria with

a key biomolecule carbamoyl phosphate (Figure 3.6) (13). While this work proposes some mechanisms where these metabolites play a role in microbial persistence, it remains to be tested whether these molecules are broadly important in helping enigmatic organism like Atribacteria survival in deep marine sediment environments.

List of References

1. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. 2012. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A* 109:16213–16216.
2. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. 2016. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J* [published online ahead of print]:doi: 10.1038/ismej.2015.233.
3. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hermsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nat Microbiol* 1:doi: 10.1038/nmicrobiol.2016.48.
4. Podar M, Makarova KS, Graham DE, Wolf YI, Koonin EV, Reysenbach A-L. 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol Direct* 8:10–1186.
5. Gupta RS, Chen WJ, Adeolu M, Chai Y. 2013. Molecular signatures for the class Coriobacteriia and its different clades; proposal for division of the class Coriobacteriia into the emended order Coriobacteriales, containing the emended family Coriobacteriaceae and Atopobiaceae fam. nov., and Eggerthellales ord. nov., containing the family Eggerthellaceae fam. nov. *Int J Syst Evol Microbiol* 63:3379–3397.
6. Hoshino T, Toki T, Ijiri A, Morono Y, Machiyama H, Ashi J, Okamura K, Inagaki F. 2017. Atribacteria from the Subseafloor Sedimentary Biosphere Disperse to the Hydrosphere through Submarine Mud Volcanoes. *Front Microbiol* 8: Article 1135.
7. Wu H, Fang Y, Yu J, Zhang Z. 2014. The quest for a unified view of bacterial land colonization. *ISME J* 8:1358–1369.
8. Doroshenko V, Airich L, Vitushkina M, Kolokolova A, Livshits V, Mashko S. 2007. YddG from *Escherichia coli* promotes export of aromatic amino acids. *FEMS Microbiol Lett* 275:312–8.
9. Kempes CP, van Bodegom PM, Wolpert D, Libby E, Amend J, Hoehler T. 2017. Drivers of Bacterial Maintenance and Minimal Energy Requirements. *Front Microbiol* 8: Article 31.

10. Paine R t. 1995. A Conversation on Refining the Concept of Keystone Species. *Conserv Biol* 9:962–964.
11. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD. 2006. Conservation of the metabolomic response to starvation across two divergent microbes. *Proc Natl Acad Sci* 103:19302–19307.
12. Kyryakov P, Beach A, Richard VR, Burstein MT, Leonov A, Levy S, Titorenko VI. 2012. Caloric restriction extends yeast chronological lifespan by altering a pattern of age-related changes in trehalose concentration. *Front Physiol* 3: Article 256.
13. Calvert S, Tacutu R, Sharifi S, Teixeira R, Ghosh P, de Magalhães JP. 2016. A network pharmacology approach reveals new candidate caloric restriction mimetics in *C. elegans*. *Aging Cell* 15:256–266.

Vita

Jordan Toby Bird was born in Fort Oglethorpe, GA. He spent his childhood in rural community of Woodlawn, AR. He completed his undergrad in Biology from the University of Central Arkansas. He joined the PhD program in UTK in the summer of 2012. He has never shied away from asking a question. Apart academic pursuits he enjoys having long conversation with his friends and brother.