7-1-2015

Diversity of understudied archaeal and bacterial populations of Yellowstone National Park: from genes to genomes

Daniel Colman

Daniel Robert Colman
Candidate

Biology
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Cristina Takacs-Vesbach , Chairperson

Robert Sinsabaugh

Laura Crossey

Diana Northup

Diversity of understudied archaeal and bacterial populations from Yellowstone National Park: from genes to genomes

by

Daniel Robert Colman

B.S. Biology, University of New Mexico, 2009

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Biology

The University of New Mexico
Albuquerque, New Mexico

July 2015

DEDICATION

I would like to dedicate this dissertation to my late grandfather, Kenneth Leo Colman, associate professor of Animal Science in the Wool laboratory at Montana State University, who even very near the end of his earthly tenure, thought it pertinent to quiz my knowledge of oxidized nitrogen compounds. He was a man of great curiosity about the natural world, and to whom I owe an acknowledgement for his legacy of intellectual (and actual) wanderlust.

ACKNOWLEDGEMENTS

I acknowledge the invaluable support, mentoring and advocacy from my dissertation advisor, Dr. Cristina Takacs-Vesbach as well as my dissertation committee members and mentors: Dr. Laura Crossey, Dr. Robert Sinsabaugh and Dr. Diana Northup. I appreciate the mentoring and support from numerous other faculty and staff in the Biology department at the University of New Mexico and Dr. William Inskeep at Montana State University. I also thank former members of the Takacs-Vesbach lab and my fellow graduate students for their support, mentorship, advice and scientific inspiration throughout my dissertation work including Dr. David Van Horn, Dr. Justine Hall, Dr. Matthew Kirk, Dr. Jordan Okie, Dr. Kendra Maas, Dr. Shannon Fitzpatrick, Levi Gray, Jason Dugger and many others. I would also like to acknowledge the personal support from Beverly Marrs and my family who were instrumental in encouraging my forward progress throughout the process of completing the research described herein.

Diversity of understudied archaeal and bacterial populations of Yellowstone National Park: from genes to genomes

by

Daniel Robert Colman

B.S. Biology, University of New Mexico 2009

Ph.D, Biology, University of New Mexico, 2015

ABSTRACT

Yellowstone National Park (YNP) thermal springs have been a crucial resource for the discovery and characterization of microbial biodiversity. The use of cultivation-independent methods has documented many new phyla of uncultured Archaea and Bacteria within thermal springs. Here, I describe the phylogenetic diversity and distribution of Archaea within the YNP thermal spring ecosystem and the phylogenetic and physiologic characterization of novel, uncultured hyperthermophilic bacterial lineages from metagenomic data.

In chapter two, I evaluated the efficacy of commonly used, 'universal' archaeal-specific 16S rRNA gene PCR primers in detecting archaeal phylogenetic diversity. In chapter three, using the PCR primers that would provide the best representation of archaeal communities, I used high-throughput 454 pyrosequencing to analyze the phylogenetic diversity and distribution of Archaea and Bacteria among 33 YNP springs. The results indicated that Archaea were ubiquitously distributed across YNP springs and exhibited significant taxonomic diversity across springs but were overall less

phylogenetically diverse in the YNP system than Bacteria. pH, followed by temperature primarily explained the distribution of both archaeal and bacterial taxonomic distribution. Co-occurence analysis suggested a substantial number of putative interactions across the YNP system between and within domains. The results from these two chapters provide the largest survey of Archaea in any thermal system to date and contribute to our understanding of their phylogenetic diversity and ecology in such systems.

In Chapter 4, I report the phylogenetic and physiologic characterization of novel, deep-branching bacterial phylotypes from metagenomic data from two YNP springs. Genome assemblies representing four populations were recovered from Aquificaceae-dominated community metagenomes from two high-temperature, circumneutral YNP springs. Phylogenetic analyses indicated they belonged to two distinct, deep-branching bacterial lineages, one of which has no currently characterized genome references. The lineages appeared to be heteroorganotrophs based on metabolic reconstructions and also were both putatively capable of using energy conserved from organic carbon degradation to fuel aerobic respiration. Analysis of the ecological distribution of these populations confirmed that they are currently restricted to high-temperature circumneutral terrestrial springs, largely within YNP. The characterization of these populations provides important physiologic context for the deepest-branching bacterial lineages and valuable genomic references for uncultured, ubiquitously distributed hyperthermophilic Bacteria.

Table of Contents

Chapter 1

Microbial biodiversity and ecology of Bacteria and Archaea that inhabit thermal springs

Context for the Study of Microbial Diversity in Thermal Springs

Thermal spring ecosystems have been an integral setting for the characterization and discovery of microbial diversity. While early cultivation-dependent work on microbial populations suggested species-depauperate microbial communities in these systems (Ward *et al.* 1998), it has become evident through cultivation-independent analyses that thermal spring communities can be exceptionally rich at multiple scales of taxonomic diversity (Barns *et al.* 1994, Hugenholtz *et al.* 1998, Mitchell 2009, Ward *et al.* 2006). The concomitant considerable diversity in physicochemical niches (e.g. pH, temperature, redox state, geologic processes contributing to the chemistry of springs) of thermal spring environments likely contributes to the broad diversity of microbial organisms present. As the global extent of microbial phylogenetic diversity is becoming more resolved, recent advances in environmental genomics have begun to provide a framework to begin connecting phylogenetic diversity with physiologic potential. In my dissertation work, I have sought to 1) determine the extent of archaeal diversity in Yellowstone National Park (YNP) hot springs using high-throughput 16S rRNA gene sequencing, 2) describe archaeal biodiversity in context of the co-occurring bacterial populations and abiotic parameters and 3) use environmental genomics to characterize the physiologic potential of novel, deeply-branching bacterial populations that are ubiquitously distributed among YNP thermal springs. In this chapter, I will first provide a background for microbial ecological research in the YNP thermal ecosystem.

The Study of Biodiversity in YNP

Early work in YNP thermal springs by Thomas Brock and others was integral in understanding the ecology of these environments in addition to the characterization of many thermophiles that have since become model organisms in thermophilic microbiology studies (e.g. *Thermus aquaticus* and *Sulfolobus acidocaldarius*; Brock and Freeze 1969, Brock *et al.* 1972). Although Brock *et al.*'s cultivation-dependent approach (and experimental observations) were foundational, later work using cultivation-independent genetic techniques (e.g. phylogenetic diversity of 16S rRNA genes) showed a considerably higher diversity of organisms in these environments than was originally thought to be present. The contrast is due to highly selective enrichment of certain organisms during cultivation, cryptic diversity and the recalcitrance of many organisms to cultivation (Ward *et al.* 1998). The application of cultivation-independent techniques revealed a significant amount of previously unknown diversity including several novel uncultivated divisions of Bacteria and Archaea that had not been previously discovered (Barns *et al.* 1994, Hugenholtz *et al.* 1998, Reysenbach *et al.* 1994). Cultivation techniques have also improved since the advent of genetic-techniques and some of these 'uncultured populations' have since been brought into laboratory culture (de la Torre *et al.* 2008, Elkins *et al.* 2008). However, due to the recalcitrance of most extremophiles to cultivation, many of the newly discovered lineages still remain uncultivated and only known from environmental surveys of 16S rRNA genes (Rinke *et al.* 2013).

Cultivation-independent techniques have also provided a relatively unbiased means to analyze the community ecology of thermal spring microbial populations. For example, the role of geochemical parameters in structuring chemosynthetic communities

has received considerable attention (Macur *et al.* 2004, Meyer-Dombard *et al.* 2005, Spear *et al.* 2005). Other studies along temperature gradients of one or a few springs have sought to explore the differentiation of populations along abiotic gradients (Hall *et al.* 2008, Meyer-Dombard *et al.* 2011, Miller *et al.* 2009). However, the lack of sampling and replication across large physicochemical gradients has often hindered the aggregation of results in a statistically rigorous manner. The most complete census of bacterial community composition in YNP thermal springs was conducted on 103 springs across the range of pH, temperature and geochemistry found in YNP springs (Mitchell 2009). The major findings of this study were that a considerable amount of bacterial diversity (including most major bacterial lineages and several novel lineages) is present within the YNP system and that pH and temperature differences were coincident with the dominance of five different bacterial community 'types' among YNP springs. That pH strongly structures microbial communities across thermal systems is consistent with other studies that confirmed this result with hydrogenase distribution across YNP (Boyd *et al.* 2010), metagenomes of 20 YNP springs (Inskeep *et al.* 2013b), archaeal communities from 22 YNP springs (Boyd *et al.* 2013), and Archaea and Bacteria from 16 springs from the Tibet/Yunnan regions of China (Song *et al.* 2013).

Archaeal Biodiversity in Thermal Springs

In contrast to the Bacteria, there is much less known about the diversity and distribution of Archaea in YNP and other thermal systems. This may be partially due to the general underdominance of Archaea in many YNP thermal springs, and the widely held hypothesis that they are only dominant in the extreme margins of thermal springs (e.g. at the highest temperatures or in acidic springs; Hugenholtz *et al.* 1998, Inskeep *et al.*

2010, Reysenbach and Shock 2002, Ward *et al.* 1998). However, recent studies document their presence and in some instances, high diversity, in springs where they were not detected previously or in springs with physicochemical profiles where they were presumed to not be present (Bowen De Leon *et al.* 2013, Boyd *et al.* 2013, Colman *et al.* 2015). Although historical sampling bias may hinder a more robust understanding of archaeal biodiversity across thermal springs in general, artifacts arising from PCR amplification bias of archaeal phylotypes may have also caused an underestimation of their biodiversity and distribution in thermal springs (Colman *et al.* 2015, Klindworth *et al.* 2013). In Chapter 2, I have analyzed archaeal-specific, 'universal' PCR primers and their accuracy in capturing the 16S rRNA gene phylogenetic diversity known to be present in a high-temperature YNP spring where metagenomic data was present to reference against PCR-based results. In Chapter 3 of this dissertation, I used the PCR primers identified as most likely to capture the greatest archaeal phylogenetic diversity and surveyed the diversity and distribution of Archaea in 33 YNP thermal springs and compared their diversity to abiotic parameters and bacterial 16S rRNA gene diversity.

Insight into Functional Diversity of Uncultured Populations from Environmental Genomics

Surveys of phylogenetic diversity provide a valuable resource to understand the community ecology of uncultured populations and baseline data to probe for evolutionarily important or industrially useful microorganisms. However, phylogenetic data only provide a means to objectively compare community composition and to discern taxonomic relationships and thus does not necessarily provide functional information that can be used to connect *in situ* function and ecology to compositional differences. Recent

advances in environmental genomics (e.g. metagenomics and single-cell genome sequencing) have begun to provide functional data for uncultured populations as well as genome references for organisms previously only known from 16S rRNA gene surveys (Baker *et al.* 2010, Dodsworth *et al.* 2013, Hedlund *et al.* 2014, Inskeep *et al.* 2013b, Kozubal *et al.* 2013, Nunoura *et al.* 2011, Rinke *et al.* 2013, Takami *et al.* 2012). Despite the increased resolution of microbial functional diversity, there are still a significant number of phylogenetic lineages without representative genomes (Rinke *et al.* 2013). These taxonomic 'gaps' hinder the accurate annotation of mixed microbial metagenomes and the study of microbial evolution at a broad evolutionary scale. Most microbial lineages, and particularly extremophiles, remain without cultivated members and/or genomic references that can be used to determine metabolic potential (Hedlund *et al.* 2014, Rinke *et al.* 2013, Wu *et al.* 2009). In a recent analysis of high-temperature, circumneutral, *Aquificeae*-dominated 'streamer' communities from YNP (Takacs-Vesbach *et al.* 2013), multiple abundant microbial populations were present but lacked homology to any close, currently characterized genomic references. In Chapter 4 of this dissertation, I characterized 1) the phylogenetic placement of these Bacteria, 2) their metabolic potential and 3) their ecological distribution among thermal springs in YNP. The populations are discussed in terms of their evolutionary relevance within Bacteria and potential *in situ* function within 'streamer' communities that are distributed globally in high temperature circumneutral environments.

References

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci U S A* **107:** 8806-8811.

Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci U S A* **91:** 1609-1613.

Bowen De Leon K, Gerlach R, Peyton BM, Fields MW (2013). Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Front Microbiol* **4:** 330.

Boyd ES, Hamilton TL, Spear JR, Lavin M, Peters JW (2010). [FeFe]-hydrogenase in Yellowstone National Park: evidence for dispersal limitation and phylogenetic niche conservatism. *ISME J* **4:** 1485-1495.

Boyd ES, Hamilton TL, Wang J, He L, Zhang CL (2013). The role of tetraether lipid composition in the adaptation of thermophilic archaea to acidity. *Front Microbiol* **4:** 62.

Brock TD, Freeze H (1969). Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. *J Bacteriol* **98:** 289-297.

Brock TD, Brock KM, Belly RT, Weiss RL (1972). *Sulfolobus*: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Archiv fur Mikrobiologie* **84:** 54-68.

Colman DR, Thomas R, Maas KR, Takacs-Vesbach CD (2015). Detection and analysis of elusive members of a novel and diverse archaeal community within a thermal spring streamer consortium. *Extremophiles* **19:** 307-315.

de la Torre JR, Walker CB, Ingalls AE, Konneke M, Stahl DA (2008). Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* **10:** 810-818.

Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG *et al* (2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* **4:** 1854.

Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, Randau L *et al* (2008). A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci U S A* **105:** 8102-8107.

Hall JR, Mitchell KR, Jackson-Weaver O, Kooser AS, Cron BR, Crossey LJ *et al* (2008). Molecular characterization of the diversity and distribution of a thermal spring microbial community by using rRNA and metabolic genes. *Appl Environ Microbiol* **74:** 4910-4922.

Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles* **18:** 865-875.

Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* **180:** 366-376.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH *et al* (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5:** e9773.

Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB, YNP Metagenome Project Steering Committee *et al* (2013). The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. *Front Microbiol* **4:** 67.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M *et al* (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41:** e1.

Kozubal MA, Romine M, Jennings R, Jay ZJ, Tringe SG, Rusch DB *et al* (2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J* **7:** 622-634.

Macur RE, Langner HW, Kocar BD, Inskeep WP (2004). Linking geochemical processes with microbial community analysis: successional dynamics in an arsenc-rich, acid-sulphate-chloride geothermal spring. *Geobiology* **2:** 163-177.

Meyer-Dombard DR, Shock EL, Amend JP (2005). Archaeal and bacterial communities in geochemically diverse hot springs of Yellowstone National Park, USA. *Geobiology* **3:** 211-227.

Meyer-Dombard DR, Swingley W, Raymond J, Havig J, Shock EL, Summons RE (2011). Hydrothermal ecotones and streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. *Environ Microbiol* **13:** 2216-2231.

Miller SR, Strong AL, Jones KL, Ungerer MC (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. *Appl Environ Microbiol* **75:** 4565-4572.

Mitchell KR (2009). Controls on microbial community structure in thermal environments; exploring Bacterial diversity and the relative influence of geochemistry and geography. Ph.D. thesis, University of New Mexico, Albuquerque.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al* (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39:** 3204-3223.

Reysenbach AL, Wickham GS, Pace NR (1994). Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60:** 2113-2119.

Reysenbach AL, Shock E (2002). Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296:** 1077-1082.

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

Song ZQ, Wang FP, Zhi XY, Chen JQ, Zhou EM, Liang F *et al* (2013). Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environ Microbiol* **15:** 1160-1175.

Spear JR, Walker JJ, McCollom TM, Pace NR (2005). Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci U S A* **102:** 2555-2560.

Takacs-Vesbach C, Inskeep WP, Jay ZJ, Herrgard MJ, Rusch DB, Tringe SG *et al* (2013). Metagenome Sequence Analysis of Filamentous Microbial Communities Obtained from Geochemically Distinct Geothermal Channels Reveals Specialization of Three Aquificales Lineages. *Front Microbiol* **4**.

Takami H, Noguchi H, Takaki Y, Uchiyama I, Toyoda A, Nishi S *et al* (2012). A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. *PLoS One* **7:** e30559.

Ward DM, Ferris MJ, Nold SC, Bateson MM (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62:** 1353-1370.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A *et al* (2006). Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units linking microbial community composition, structure and function. *Philos T R Soc B* **361:** 1997-2008.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN *et al* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462:** 1056-1060.

Chapter 2

Detection and Analysis of Elusive Members of a Novel and Diverse Archaeal Community Within a Thermal Spring Streamer Consortium

Daniel R. Colman[1], Raquela Thomas[2], Kendra R. Maas[3], Cristina D. Takacs-Vesbach[1]

[1]Department of Biology, University of New Mexico, Albuquerque, NM, USA

[2]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston, SC, USA; [3]Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada

Citation: *Extremophiles* 2015 v. 19, No. 2, pp. 307-315

Abstract

Recent metagenomic analyses of Yellowstone National Park (YNP) thermal spring communities suggested the presence of minor archaeal populations that simultaneous PCR based assays using traditional 'universal' 16S rRNA gene primers failed to detect. Here we use metagenomics to identify PCR primers effective at detecting elusive members of the Archaea, assess their efficacy, and describe the diverse and novel archaeal community from a circum-neutral thermal spring from the Bechler region of YNP. We determined that a less commonly used PCR primer, Arch349F, captured more diversity in this spring than the widely used A21F primer. A search of the PCR primers against the RDP 16S rRNA gene database indicated that Arch349F also captured the largest percentage of Archaea, including 41% more than A21F. Pyrosequencing using the Arch349F primer recovered all of the phylotypes present in the clone-based portion of the study and the metagenome of this spring in addition to several other populations of Archaea, some of which are phylogenetically novel. In contrast to the lack of

amplification with traditional 16S rRNA gene primers, our comprehensive analyses suggested a diverse archaeal community in the Bechler spring, with implications for recently discovered groups such as the Geoarchaeota and other undescribed archaeal groups.

Introduction

Hydrothermal vent ecosystems have been a rich resource for the study of microbial diversity, astrobiology and geomicrobiology for several decades. Much of the diversity in these systems is only known from environmental genomics, due to the recalcitrance of most thermal spring organisms to culturing. In particular, the study of archaeal biodiversity in hydrothermal systems has been driven by environmental genotyping, and now environmental genomics. Recent evidence, however, suggests there remains a methodological gap between the ability to detect Archaea and our understanding of their ubiquity in nature. Although universal 16S rRNA gene primers such as 21F/1492R (DeLong 1992) have lead to the discovery of many new lineages (Barns *et al.* 1994, DeLong 1992, Takai and Sako 1999) and transformed our view of archaeal phylogenetic diversity during the last 20 years (Barns *et al.* 1994, Hugenholtz 2002, Schleper *et al.* 2005), metagenomic studies highlight the fact that these primers are not 'universal'. In a recent metagenomic survey of twenty Yellowstone National Park thermal springs (Inskeep *et al.* 2013b), 16S rRNA gene clone-based surveys using 21F/1492R failed to detect Archaea in seven of the samples (Table 2.S1). The seven springs lacking Archaea by PCR panel were all of circum-neutral to alkaline pH (6.2-9.1) and all contain minor (< 1.0% of metagenomic reads), and in some cases greater populations of Archaea based on classification of metagenomic reads (For example,

springs from the Calcite Springs and Bechler regions; Inskeep *et al.* 2010, Takacs-Vesbach *et al.* 2013; Table 2.S1). The disparity in apparent community composition between PCR based studies using 'universal' 16S rRNA gene primers and a less taxonomically biased approach, such as metagenomic analysis, suggests an inability of traditional primers to accurately represent archaeal community composition. Inaccuracy in community composition analyses, even of minor populations, can hamper efforts to discover novel biodiversity critical for evolutionary studies as well as understanding distributions of known taxa. For example, current efforts to genomically characterize low-population organisms that may be evolutionary important from single cell genomics rely on accurate and complete community censuses to find genomic targets (Rinke *et al.* 2013).

Much initial archaeal biodiversity survey work was focused on placing novel lineages into robust phylogenetic context (Barns *et al.* 1994, Takai and Sako 1999), which necessitated the largest gene fragment possible, such as produced by the 21F/1492R primer combination (Olsen *et al.* 1986). However, with the advent of next-generation sequencing technologies, focus has shifted to high-throughput, accurate estimations of taxonomic diversity which do not require full-length gene fragments (Liu *et al.* 2007) and allows researchers to describe low abundance populations that may be ecologically relevant or important for evolutionary and bioprospecting studies. Many effective archaeal primers have been reported previously (for review see: Baker *et al.* 2003, Klindworth *et al.* 2013), but a robust analysis of their efficiency has not been evaluated in samples where there exists communities that were not detected by the more commonly used 21F/1492R primer combination. Further, taxonomic biases of PCR

primers in the amplification of prokaryotic 16S rRNA genes in mixed communities have been previously documented in many systems (For example: Baker *et al.* 2003, Klindworth *et al.* 2013, Kumar *et al.* 2011, Suzuki and Giovannoni 1996). However, there has been little empirical testing of archaeal-specific PCR primers against a relatively unbiased reference, such as a metagenomic, shotgun-sequence produced dataset for a given sample. Here we tested universal archaeal 16S rRNA gene PCR primer combinations by comparing the archaeal communities described metagenomically here and in Takacs-Vesbach et al. (2013) to population results derived from complementary *in silico*, clone library and 454 pyrosequencing approaches.

Materials and Methods

The Bechler region spring is a high temperature (~80 ^{o}C), alkaline spring (pH 7.8) that contains dense *Thermocrinis* spp. filament streamers along runoff channels. Archaea are also associated with the *Thermocrinis* filaments (Takacs-Vesbach *et al.* 2013). Protein-coding genes belonging to Archaea in the publically available spring metagenome (based on $\geq$ 60% BLAST identity classifications) were analyzed using the IMG/M web interface (IMG metagenome taxon ID: 2013515002; Markowitz *et al.* 2012) to examine community diversity. The taxonomic affiliations of 16S rRNA genes present in the metagenomic data were assessed by phylogenetic analysis of partial 16S rRNA genes that were present for two phylotypes. 16S rRNA gene references were chosen by top uncultured clone BLAST hits in Genbank and supplemented with references for the major phylogenetic lineages of Crenarchaeota. 16S rRNA genes were aligned using pyNAST (Caporaso *et al.* 2010a) and trimmed so that all of the sequences shared the same homologous 691 positions common to the two metagenomic 16S rRNA gene

fragments. Phylogenetic structure was explored using the MEGA software package (Tamura *et al.* 2013), and a final maximum likelihood tree was constructed using the GTR + G + I evolutionary model as suggested by jModelTest 2 (Darriba *et al.* 2012). The geochemical context of the Bechler spring and detailed microbial metagenomic analyses are described in Takacs-Vesbach et al. (2013).

Archaeal 16S rRNA genes were amplified from the Bechler spring using 11 previously published archaeal-specific primers (Table 2.S2) with the environmental DNA used in the metagenomic analysis and in the 21F/1492R 16S rRNA gene screens reported in Takacs-Vesbach et al. (2013). PCR reactions consisted of 10 µl 5X GoTaq buffer (Promega, Madison, WI, USA), 12.5 mM dNTP (BioLine USA Inc., Taunton, MA, USA), 20 pmol of both primers, 2.5 units of GoTaq DNA polymerase (Promega), 1 µl of 2% (w/v) bovine serum albumin, 1 mM $MgCl_2$, and ~50 ng DNA. The PCR cycling program consisted of 30s at 94^{o}C, 30s at 55^{o}C or 58^{o}C (optimized for different primer pairs) and 90s at 72^{o}C for a total of 30 cycles and was performed on an ABI GeneAmp 2700 (Life Technologies, Grand Island, NY, USA). Each forward primer was tested with each reverse primer, except 571F, which was tested with 1391R only, but failed to amplify any products. Successful amplifications were duplicated, combined and gel purified with a Wizard SV gel purification kit (Promega). PCR products were cloned with the pGEM-T Easy kit (Promega). Between 13 and 18 cloned inserts for each primer set were sequenced using the BigDye terminator cycle sequencing kit (Life Technologies) with the M13F primer on an ABI 3130x genetic analyzer (Life Technologies). Primer sets using the 1391R reverse primer produced larger bands consistent with *Pyrobaculum* sp. containing 16S rRNA gene introns (Itoh *et al.* 2003), which were gel excised and up to 8

clones were sequenced to confirm the presence/absence of *Pyrobaculum* sp. Clone insert identity was evaluated by aligning the cloned 16S rRNA gene sequences in mothur (Schloss *et al.* 2009), manually curating the alignment and measuring phylogenetic distances against reference 16S rRNA gene sequences from the metagenomic dataset and closely-related Genbank clones.

Potential archaeal diversity detected by the primers was also measured by querying the RDP database of archaeal 16S rRNA genes using the Probe Match function (http://rdp.cme.msu.edu/probematch/search.jsp; Cole *et al.* 2014). Queries against RDP were conducted only on 16S rRNA gene sequences containing the *Escherichia coli* base positions encompassed by all primers (8-585 for the forward primers and 787-1407 for the reverse primers) and with 0 mismatched bases allowed. To incorporate a 16S rRNA gene dataset produced without the potential biases inherent in PCR-derived data, the primers were also tested against archaeal 16S rRNA genes present in the YNP community sequencing project described in Inskeep et al. (2013). A total of 76 archaeal 16S rRNA gene sequences from the 20 metagenomes were downloaded from the IMG/M database. The metagenome-derived 16S rRNA gene sequences were tested against each primer individually to avoid biases from comparisons to incomplete 16S rRNA gene sequences in the metagenomic contigs.

To probe the full extent of archaeal diversity in the sample, the two forward primers with the largest RDP database coverage, Arch349F and 340F, and the reverse primer A915R were used in 454 pyrosequencing analysis of the spring's archaeal 16S rRNA gene diversity. Barcoded amplicon pyrosequencing was conducted as described previously (Andreotti *et al.* 2011, Dowd *et al.* 2008). Briefly, 100 ng of DNA per sample

was amplified in triplicate by a single step PCR to create 16S rRNA gene amplicons containing the Roche-specific sequencing adapters (454 Life Sciences, Branford, CT, USA) and a barcode unique to each sample. Amplicons were purified using Agencourt Ampure beads and combined in equimolar concentrations. Pyrosequencing was performed on a Roche 454 FLX instrument using Roche titanium reagents and titanium procedures. The 16S rRNA gene sequences were quality filtered, denoised, screened for PCR errors, and chimera checked using AmpliconNoise and Perseus to minimize potential amplification and sequencing artifacts (Quince *et al.* 2011). Denoised reads were classified using the Quantitative Insights Into Microbial Ecology (QIIME) package pipeline after clustering into OTUs and picking an OTU representative to assign taxonomic classification to (Caporaso *et al.* 2010b). OTUs were compared against references for the phylotypes found in the clone library and metagenomic datasets using a local blastn search. Raw 454 pyrosequencing data from this study are available through the NCBI Sequence Read Archive as SRP049556. The individual sff files from this study were assigned the accession numbers SAMN03145649 through SAMN03145650 under Bioproject PRJNA265119. Clone library-produced 16S rRNA gene sequence data (accessions KP091542 - KP091669) are available through Genbank.

Results and Discussion

The Bechler spring metagenome contained a minor population of Archaea (8% of the total metagenomic reads). Taxonomic classification of the assembled metagenome archaeal reads indicated a predominant Thermoproteaceae population (92% of all archaeal reads), primarily in the genus *Pyrobaculum* (99% of Thermoproteaceae reads). However, higher diversity was suggested by the classification of the remaining genes into

17 other archaeal families, including minor populations (< 5 % of archaeal reads) binned in families of the phyla Euryarchaeota, Thumarchaeota and Korarchaeota. Only two archaeal 16S rRNA gene sequence phylotypes were present in the metagenomic dataset. One phylotype was closely related to *Pyrobaculum* spp. (IMG gene id YNP13_01430) with close relation to *Pyrobaculum neutrophilum* strain V24Sta (99% sequence identity by BLAST, Figure 2.1). The second phylotype (IMG gene ID YNP13_216600) was only closely related to a single uncultured clone from Great Boiling Spring, Great Basin, Nevada (GBS_L2_E12, Figure 2.1). Both of these uncultured phylotypes were monophyletic with 16S rRNA genes from the recently described NAG1 genome (IMG taxon id 2504756013) in the proposed Geoarchaeota phylum (Kozubal *et al.* 2013), in addition to uncultured clones associated with Geoarchaeota from YNP thermal springs (Kozubal *et al.* 2012) and a phylotype closely related to the Geoarchaeota from a seafloor vent biofilm near Papua, New Guinea (clone PNG_TBR_A43; Meyer-Dombard *et al.* 2013). The phylum-level designation of the Geoarchaeota has been debated, as various gene combinations provide contradictory phylogenetic placement of the group within and separate from the Crenarchaeota phylum (Guy *et al.* 2014). Our overall archaeal topology is consistent with Kozubal et al. (2013), but differs from the sister-level relationship between Thermoproteales and Geoarchaeota using more robust phylogenetic analyses (Guy *et al.* 2014). Regardless, the intragroup clades were all highly supported for the Geoarchaeota. The Geoarchaeota-affiliated clones and NAG1 genome were all detected at acidic, Fe-rich YNP thermal springs (Kozubal *et al.* 2012, Kozubal *et al.* 2013) and the single seafloor vent biofilm clone was also detected in a slightly acidic, Fe-rich environment (Meyer-Dombard *et al.* 2012, Meyer-Dombard *et al.* 2013). However, both

the Bechler region spring and Great Basin spring are alkaline and do not contain appreciable levels of Fe (Costa *et al.* 2009, Dodsworth *et al.* 2011, Takacs-Vesbach *et al.* 2013), suggesting that the phylotype detected here may constitute an ecologically and phylogenetically distinct second clade within the Geoarchaeota.

All but one PCR primer set amplified products for the clone-based primer comparison component of the study, but only two were able to detect both archaeal 16S rRNA gene phylotypes present in the metagenomic data. A third phylotype, whose 16S rRNA gene was not recovered by the modest metagenomic sequencing effort in Takacs-Vesbach et al. (2013), was affiliated with the Aigarchaeota group of Archaea (represented by Genbank accession HM448082) and detected by many of the primer sets. All clones shared >97% identity to one of seven phylotypes (Table 2.1). Both of the primer sets that amplified the three predominant phylotypes used the forward primer Arch349F. The *in silico* primer comparison results emphasized that the forward primers Arch21F, A109F and A571F match sequences that, as a whole, are subsets of those matched with Arch349F and A340F, which both had the highest coverage of RDP and IMG records (Table 2.1, Figure 2.2). This is consistent with a recent *in silico* analysis of archaeal primers which indicated that Arch349F had the highest overall coverage of database deposited archaeal 16S rRNA gene sequences (Klindworth *et al.* 2013). Reverse primers did not differ considerably in record matches (89 - 92% RDP records matched by each).

Pyrosequencing of archaeal 16S rRNA genes revealed additional novel archaeal diversity, although taxonomic differences were observed between the two primer sets that were used (Figure 2.3). A340F primarily amplified unidentified taxa within the candidate Parvarchaeota phylum (identified as a the proposed order 'Micrarchaeles' in the latest

Greengenes taxonomy), and a minor percentage of Aigarchaeota phylotypes. The Parvarchaeota are known primarily from acidic environments such as acid-mine drainage biofilms (Baker *et al.* 2010, Rinke *et al.* 2013), so it's unlikely they are a major, active component of the Bechler community, especially considering that Crenarchaeota were the dominant phylum in the metagenomic dataset. Further, neither the *Pyrobaculum* sp., nor Geoarchaeota-like phylotypes that were present in the metagenomic dataset were detected by A340F, which suggests potential amplification biases. In contrast, 454 analyses with Arch349F detected all three phylotypes found in the metagenomic and clone library data in addition to several other populations. The Arch349F results are generally consistent with the metagenomic and clone library data, although the Caldiarchaceae (proposed family of the Aigarchaeota in the latest Greengenes taxonomy) populations were overrepresented relative to *Pyrobaculum* populations. Arch349F also detected a larger relative percentage of organisms that were not classifiable at the phylum level in addition to minor populations of unclassified Thermoprotei, *Nitrosocaldus* sp., Desulfurococcaceae-affiliated organisms and *Caldiarchaeum* spp. not detected by A340F. Both primer sets amplified a small percentage of bacterial reads primarily classified as the family Aquificaceae, which contains the predominant microbial organism present in this community, *Thermocrinis* sp.(Takacs-Vesbach *et al.* 2013). While Arch349F only matches 1.3×10^{-5} % of RDP bacterial records (35 total of mostly unclassified bacteria), stringency could be increased with minimal loss of matched archaeal sequences by substituting a G for the K and A for the W in the 10th and 17th residues, respectively (0/80% and 2.7×10^{-6}/94% bacterial/archaeal records matched in RDP with 0 and 1 mismatched bases, respectively).

Our results suggest that commonly used primers may not be adequate for describing the archaeal diversity in thermal spring systems, and primers such as Arch349F would more accurately reflect the *in situ* archaeal community. Pyrosequencing with Arch349F indicated the archaeal community in this spring is taxonomically diverse, which contrasts with previous indications by unsuccessful PCR amplifications using traditional 16S rRNA gene primers and metagenomic analyses that only describe predominant taxonomic populations. While all phylotypes that were recovered by clone library and metagenomic analyses were also recovered using the forward primer Arch349F, several populations that couldn't be classified at or below phylum-level taxonomy were also recovered. Further, the nearly complete matching of the RDP database by Arch349F suggests that our results would likely be effective in other ecosystems. Previous database analyses of archaeal primers are concordant with our results that the Arch349F primer provides the best overall coverage of archaeal diversity and is appropriate for other short-read sequencing platforms such as illumina and Ion Torrent (Klindworth *et al.* 2013). In conclusion, our results suggest that leveraging next generation sequencing and PCR primers with broader archaeal specificity provide not only greater accuracy in community composition analyses, but may aid in the detection of novel, minor populations of Archaea from environmental samples.

Acknowledgements

This work was funded by National Science Foundation (NSF) DEB0206773 and the Department of Energy-Joint Genome Institute Community Sequencing Program (CSP 787081). DC received additional support from NSF EHR0832947, RT from National Institutes of Health (NIH) R25 GM075149, and KM from the New Mexico Space Grant

Consortium. Technical support for sequencing was provided by the University of New Mexico's Molecular Biology Facility, which is supported by NIH grant P20GM103452 from the Institute Development Award Program of the National Center for Research Resources. We thank W. Inskeep for comments on the manuscript and providing PCR panel success results used in Table 2.S1.

References

Andreotti R, Pérez de León AA, Dowd SE, Guerrero FD, Bendele KG, Scoles GA (2011). Assessment of bacterial diversity in the cattle tick *Rhipicephalus* (*Boophilus*) *microplus* through tag-encoded pyrosequencing. *BMC Microbiol* **11:** 6.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci U S A* **107:** 8806-8811.

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

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26:** 266-267.

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

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y *et al* (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42:** D633-642.

Costa KC, Navarro JB, Shock EL, Zhang CLL, Soukup D, Hedlund BP (2009). Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. *Extremophiles* **13:** 447-459.

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9:** 772.

DeLong EF (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci U S A* **89:** 5685-5689.

Dodsworth JA, Hungate BA, Hedlund BP (2011). Ammonia oxidation, denitrification and dissimilatory nitrate reduction to ammonium in two US Great Basin hot springs with abundant ammonia-oxidizing archaea. *Environ Microbiol* **13:** 2371-2386.

Dowd SE, Wolcott RD, Sun Y, McKeehan T, Smith E, Rhoads D (2008). Polymicrobial nature of chronic diabetic foot ulcer biofilm infections determined using bacterial tag encoded FLX amplicon pyrosequencing (bTEFAP). *PLoS One* **3:** e3326.

Guy L, Spang A, Saw JH, Ettema TJ (2014). 'Geoarchaeote NAG1' is a deeply rooting lineage of the archaeal order Thermoproteales rather than a new phylum. *ISME J* **8:** 1353-1357.

Hugenholtz P (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3:** Reviews0003.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH *et al* (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5:** e9773.

Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB, YNP Metagenome Project Steering Committee *et al* (2013). The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. *Front Microbiol* **4:** 67.

Itoh T, Nomura N, Sako Y (2003). Distribution of 16S rRNA introns among the family Thermoproteaceae and their evolutionary implications. *Extremophiles* **7:** 229-233.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M *et al* (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41:** e1.

Kozubal MA, Macur RE, Jay ZJ, Beam JP, Malfatti SA, Tringe SG *et al* (2012). Microbial iron cycling in acidic geothermal springs of Yellowstone National Park: integrating molecular surveys, geochemical processes, and isolation of novel Fe-active microorganisms. *Front Microbiol* **3:** 109.

Kozubal MA, Romine M, Jennings R, Jay ZJ, Tringe SG, Rusch DB *et al* (2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J* **7:** 622-634.

Kumar PS, Brooker MR, Dowd SE, Camerlengo T (2011). Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *Plos One* **6**.

Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res* **35:** e120.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y *et al* (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40:** D123-129.

Meyer-Dombard DR, Price RE, Pichler T, Amend JP (2012). Prokaryotic populations in Arsenic-Rich Shallow-Sea Hydrothermal Sediments of Ambitle Island, Papua New Guinea. *Geomicrobiology J* **29:** 1-17.

Meyer-Dombard DR, Amend JP, Osburn MR (2013). Microbial diversity and potential for arsenic and iron biogeochemical cycling at an arsenic rich, shallow-sea hydrothermal vent (Tutum Bay, Papua New Guinea). *Chem Geol* **348:** 37-47.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40:** 337-365.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing Noise From Pyrosequenced Amplicons. *Bmc Bioinformatics* **12**.

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

Schleper C, Jurgens G, Jonuscheit M (2005). Genomic studies of uncultivated archaea. *Nat Rev Microbiol* **3:** 479-488.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al* (2009). Introducing mothur: Open-source, Platform-Independent, Community-supported Software for Describing and Comparing Microbial Communities. *Appl Environ Microbiol* **75:** 7537-7541.

Suzuki MT, Giovannoni SJ (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625-630.

Swingley WD, Meyer-Dombard DR, Shock EL, Alsop EB, Falenski HD, Havig JR *et al* (2012). Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem. *PLoS One* **7:** e38108.

Takacs-Vesbach C, Inskeep WP, Jay ZJ, Herrgard MJ, Rusch DB, Tringe SG *et al* (2013). Metagenome Sequence Analysis of Filamentous Microbial Communities Obtained from Geochemically Distinct Geothermal Channels Reveals Specialization of Three Aquificales Lineages. *Front Microbiol* **4**.

Takai K, Sako Y (1999). A molecular view of archaeal diversity in marine and terrestrial hot water environments. *FEMS Microbiol Ecol* **28:** 177-188.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30:** 2725-2729.

Tables

Table 2.1. Phylotypes detected by each primer set and percentage of database records matched by each set[a]

Primer Combination	YNP13_01430[b]	YNP13_216600[b]	HM448082[c]	HM448111[c]	JX316758[c]	Bacteria[c]	RDP	IMG
Arch349F/A915R	X	X	X		X		87/89	83/42
Arch349F/A806R	X						87/92	83/100
Arch349F/1391R	X	X	X			EU156156	87/92	83/67
Arch21F/A915R	X		X	X			46/89	48/42
Arch21F/Arch806R			X				46/92	48/100
Arch21F/1391R			X				46/92	48/67
A109F/A915R	X		X				63/89	31/42
A109F/Arch806R	X		X			AJ320219	63/93	31/100
A109F/1391R	X						63/92	31/67
A571F/1391R							62/92	23/67
A340F/915R[d]							83/89	93/42

[a] Positive detection of phylotype (at least one clone insert sequence < 0.03 diverged from reference 16S rRNA gene sequence) indicated by X
[b] IMG Gene ID of 16S rRNA gene phylotype detected in metagenomic analysis of sample (JGI metagenome ID: 2013515002)
[c] Accession ID given of best BLAST match with >97% nucleotide similarity
[d] Not used in clone library based analysis

Figures

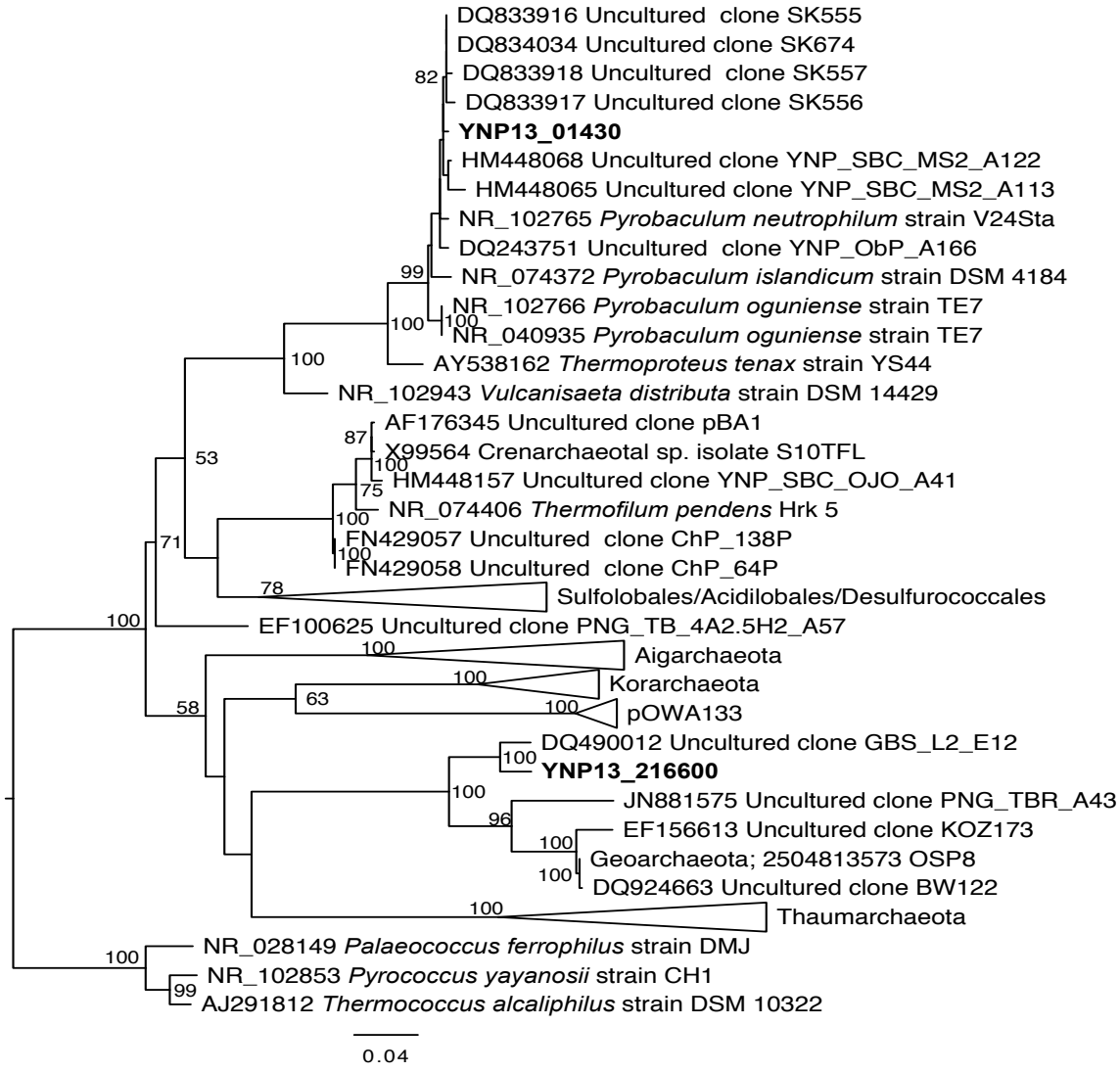


Figure 2.1 Maximum likelihood phylogenetic tree of both 16S rRNA gene sequences present in the Bechler spring metagenome. Bootstrap values are given at each node (out of 100 bootstraps) where values >50. The phylogeny is rooted with three Euryarchaeote organisms.

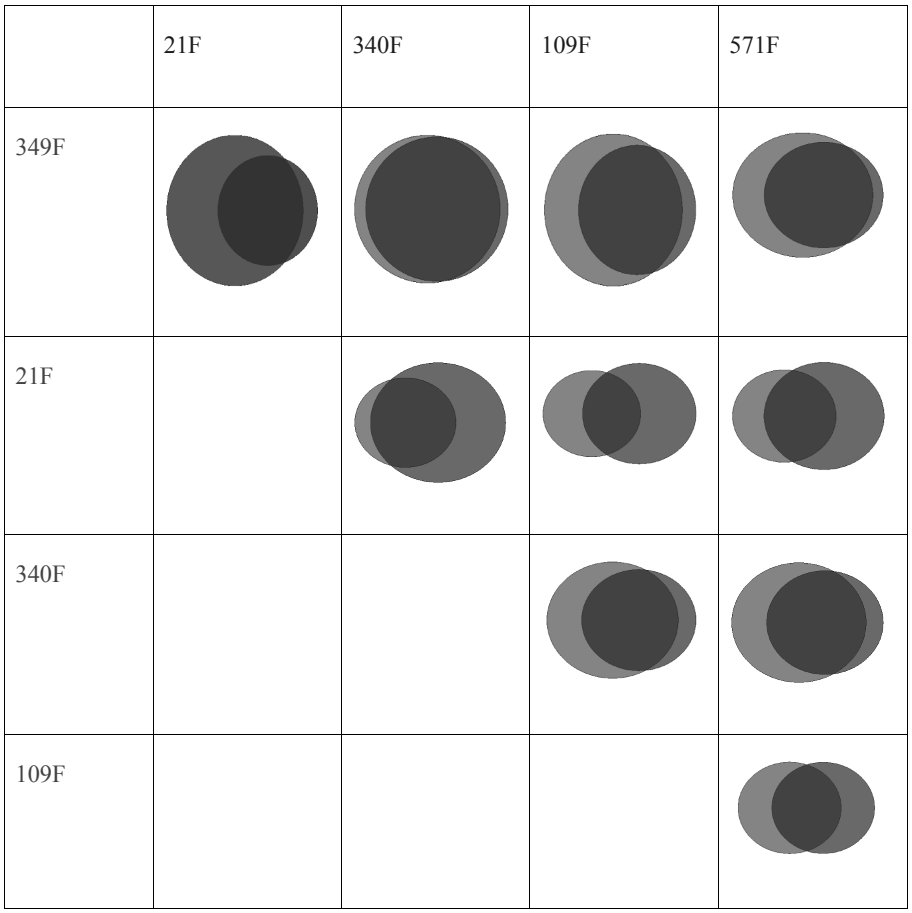


Figure 2.2 Venn diagrams of 16S rRNA gene sequence record overlap detected in the RDP database by pairwise comparisons between forward primers of this study. Each primer comparison used only the 16S rRNA gene records from the RDP archaeal 16S rRNA database that covered the total range of the forward primers (*E. coli* base positions 8-585; 18,321 16S rRNA gene records total). Circle size and overlap is proportional to total number of records matched. Primers listed in the first column are circles on the left, and those listed in the first row are circles on the right.

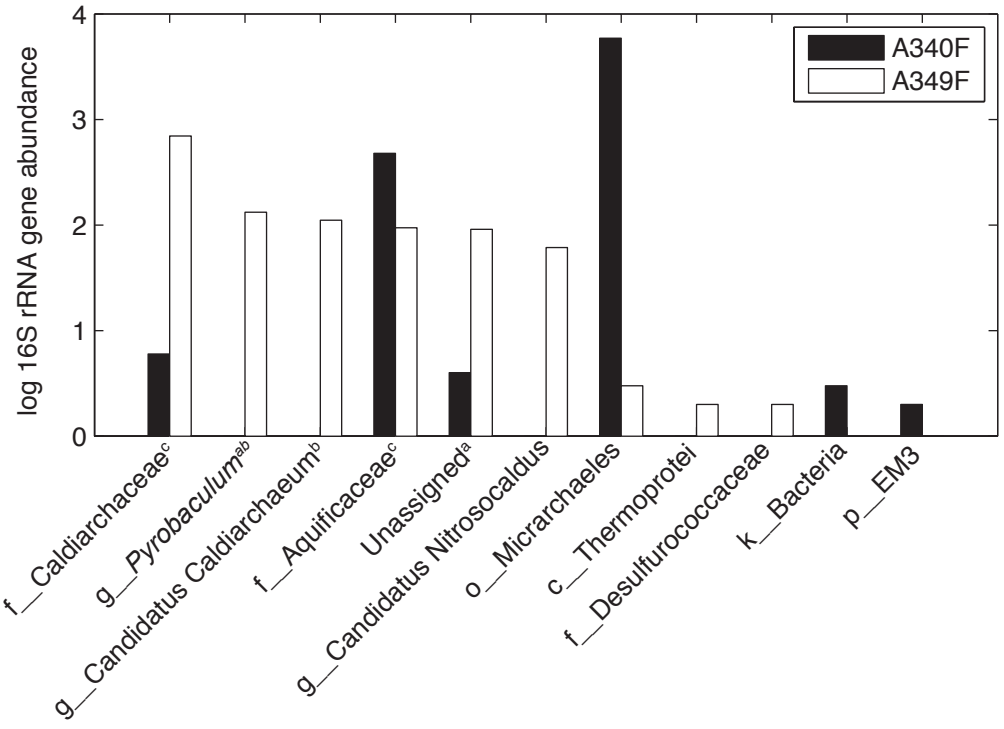


Figure 2.3 Bar chart showing the abundance and taxonomic classification of 454 Pyrosequencing reads for primer sets Arch349F-915R (read n=1190) and A340F-915R (read n=6369). Taxonomic groups are arranged in descending order by abundance in the Arch349F-915R dataset. Best classification level is given, with taxonomic hierarchy level preceding taxa name: k-kingdom, p-phyla, c-class, f-family, g-genus and s-species. [a]Taxonomic classification of a shared OTU between Arch349F and the metagenomic dataset. [b]Taxonomic classification of a shared OTU between the clone library reference dataset and Arch349F. [c]Taxonomic classification of a shared OTU between the clone library reference dataset and both A340F and Arch349F.

Supplementary Tables.

Table 2.S1. Information for PCR Panel results from metagenomic survey of 20 YNP hot springs and taxonomic composition of reads

Sample ID	Site	Archaeal 16S PCR[a]	Bacterial 16S PCR[a]
YNP_1	"Alice Springs", Crater Hills	SHTH	
YNP_2	West Nymph Lake	SHGN	SHGO
YNP_3	Monarch Geyser	SHFY	SHFZ
YNP_4	Joseph's Coat	SHGW	SHGX
YNP_5	Green Chloroflexus mats	SHIF	SHIH
YNP_6	White Creek		SHGT
YNP_7	Chocolate Pots		SHHA
YNP_8	KOZ	SHHC	SHHF
YNP_9	Dragon Spring	SHTC	SHTF
YNP_10	Narrow Gauge		SHOS
YNP_11	Octopus Spring	SHOU	SHOW
YNP_12	Calcite Springs		SHOZ
YNP_13	Bechler 50		SHGC
YNP_14	Grey Streamers	SHIA	SHIB
YNP_15	Mushroom Spring	SHYF	SHYC
YNP_16	Fairy Geyser		SHYH
YNP_17	Obsidian Pool	SHYO	SHYP
YNP_18	Washburn Springs	SIAW	SIAX
YNP_19	Cistern Spring	SIAS	
YNP_20	Purple-Sulfur		SIAO

[a] Black indicates not detected, % reads estimated from assembly with larger gene count using IMG web server

Sample ID	pH	Temp	% Reads classified as Archaea
YNP_1	2.6	76	45.03
YNP_2	4	88	37.09
YNP_3	4.4	78-80	27.87
YNP_4	6.1	80	46.16
YNP_5	6.2	56-57	0.61
YNP_6	8.2	48-50	0.16
YNP_7	6.2	52	0.16
YNP_8	3.4	72	35.21
YNP_9	3.1	68-72	12.72
YNP_10	6.5	70-72	0.44
YNP_11	7.9	80-82	8.34
YNP_12	7.8	74-76	1.6
YNP_13	7.8	80-82	8.36
YNP_14	3.5	72-74	18.8
YNP_15	8.2	60	1.09
YNP_16	9.1	36-38	0.11
YNP_17	5.7	56	1
YNP_18	6.4	76	15.55
YNP_19	4.4	78-80	46.78
YNP_20	6.2	54-56	0.59

Sample ID	% Reads unclassified	% Reads Dominant Group	Dominant Group
YNP_1	47.59	44.62	Crenarchaeota
YNP_2	58.16	36.49	Crenarchaeota
YNP_3	66.84	27.32	Crenarchaeota
YNP_4	48.77	45.88	Crenarchaeota
YNP_5	64.7	9.04	Chloroflexi
YNP_6	68.78	11.4	Cyanobacteria
YNP_7	65.69	4.49	Cyanobacteria
YNP_8	55.33	34.36	Crenarchaeota
YNP_9	78.49	10.07	Crenarchaeota
YNP_10	35.14	15.7	Aquificae
YNP_11	77.72	8.68	Aquificae
YNP_12	37.86	15.64	Aquificae
YNP_13	59.23	25.45	Aquificae
YNP_14	62.65	17.8	Crenarchaeota
YNP_15	70.52	3.72	Chloroflexi
YNP_16	68.46	7.91	Proteobacteria
YNP_17	77.36	5.62	Proteobacteria
YNP_18	70.61	12.24	Crenarchaeota
YNP_19	48.65	46.53	Crenarchaeota
YNP_20	68.05	7.5	Chloroflexi

Table 2.S2. Information for primers used in this study

Primer	Reference	Sequence (5' -> 3')	*E. coli* 16S rRNA gene bases covered
Arch349F	(Takai and Horikoshi 2000)	GYGCASCAGKCGMGAAW	349-365
A340F	(Gantner *et al.* 2011)	CCCTAYGGGGYGCASCAG	340-357
A571F	(Baker *et al.* 2003)	GCYTAAAGSRDCCGTAGC	568-585
A109F	(Groβkopf *et al.* 1998)	ACKGCTCAGTAACACGT	109-125
Arch21F	(DeLong 1992)	TTCCGGTTGATCCYGCCGGA	8-26
Arch806R	(Takai and Horikoshi 2000)	GGACTACVSGGGTATCTAAT	787-806
A915R	(DeLong 1992)	GTGCTCCCCCGCCAATTCCT	915-934
1391R	(Lane 1991)	GACGGGCGGTGWGTRCA	1391-1407

Supplemental References

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

DeLong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci U S A 89:5685-5689

Gantner S, Andersson AF, Alonso-Sáez L, Bertilsson S (2011) Novel primers for 16S rRNA-based archaeal community analyses in environmental samples. J Microbiol Meth 84:12-18

Groβkopf R, Janssen PH, Liesack W (1998) Diversity and structure of the methanogenic community in anoxic rice paddy soil microcosms as examined by cultivation and direct 16S rRNA gene sequence retrieval. Appl Environ Microb 64:960-969

Lane DJ (1991) 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds) Nucleic acid techniques in bacterial systematics. John Wiley & Sons, Chichester, United Kingdom, pp 115-175

Takai K, 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

Chapter 3

Phylogenetic Diversity and Distribution of Archaea and Bacteria Among Yellowstone National Park Thermal Springs

Colman, D.R.[1], Okie, J.[2] and Takacs-Vesbach C.D.[1]

[1]Department of Biology, University of New Mexico, Albuquerque, NM, USA; [2]School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

To be submitted to *The International Society for Microbial Ecology Journal*

Abstract

Once thought to be environmentally restricted to the extreme margins of habitability, Archaea are now known to be more phylogenetically diverse and ubiquitously distributed globally. Their diversity and distribution across the physicochemical gradients present in hydrothermal systems has however remained largely unknown. We used high-throughput 454 pyrosequencing of archaeal and bacterial 16S rRNA gene sequences from 32 Yellowstone National Park (YNP) thermal springs to assess archaeal diversity and distribution across the YNP thermal spring ecosystem. Archaea were detected in every spring, and despite a lower overall richness than Bacteria in YNP, most springs harbored similar levels of richness of the two domains. pH, followed by temperature explained most of the variation in archaeal and bacterial distribution across springs. Taxonomic lineages were distributed largely in accordance with discrete differences in spring pH and temperature profile. Network analysis suggested myriad putative interactions of OTUs across and within-domains. The newly described Aigarchaeota were a central component of mid-high pH OTU networks. Our results significantly expand our understanding of archaeal diversity and distribution in

thermal spring systems and highlight community members such as the Aigarchaeota that may be previously unappreciated and important contributors to community function in YNP springs.

Introduction

The discovery of a third domain of life, the Archaea, by Woese *et al.* was a watershed moment in microbiology (Woese and Fox 1977). Early understanding of their ecology suggested a distribution and diversity restricted to extreme environments, where many of the first known archaeal species were isolated from. However, the use of cultivation-independent techniques for studying environmental microbial populations has indicated that Archaea are ubiquitously distributed among nearly every ecosystem, and are also important in global nutrient cycling (Offre *et al.* 2013). The use of cultivation-independent techniques has also shown that archaeal phylogenetic diversity is much greater than was originally supposed (Barns *et al.* 1994, Castelle *et al.* 2015, Takai and Horikoshi 1999), with a number of new lineages only recently being characterized with the use of high-throughput environmental genomics (Baker *et al.* 2010, Castelle *et al.* 2015, Kozubal *et al.* 2013, Nunoura *et al.* 2011, Rinke *et al.* 2013, Spang *et al.* 2013, Spang *et al.* 2015). Many uncultured archaeal lineages feature prominently in the study of early evolution of all life and particularly in eukaryogenesis (Nunoura *et al.* 2011, Spang *et al.* 2015), contribute to our understanding of the physiological diversity of life (Baker *et al.* 2010), and are integral in understanding nutrient cycling (Offre *et al.* 2013). Yet, our understanding of the diversity and distribution of most archaeal taxonomic groups is lacking, including in hydrothermal ecosystems, where they may be most abundant and diverse (Auguet *et al.* 2010, Barns *et al.* 1994).

Hydrothermal spring systems have been rich resources for the study of Archaea. Environmental genetic surveys in terrestrial springs of Yellowstone National Park (YNP), Tibet, China, the United States Great Basin, Iceland, and Russian thermal fields have indicated a broad and diverse global distribution of Archaea among thermal springs (Auchtung *et al.* 2006, Auchtung *et al.* 2011, Barns *et al.* 1994, Bowen De Leon *et al.* 2013, Boyd *et al.* 2013, Costa *et al.* 2009, Huang *et al.* 2011, Inskeep *et al.* 2005, Inskeep *et al.* 2013a, Kozubal *et al.* 2012, Macur *et al.* 2004, Meyer-Dombard *et al.* 2011, Miller-Coleman *et al.* 2012, Perevalova *et al.* 2008, Reigstad *et al.* 2010, Song *et al.* 2013, Wang *et al.* 2013). However, the analysis of a single taxonomic group, one or a few springs of a single thermal field or the lack of sampling across relevant physicochemical parameters often hinders the aggregation of results across studies to mechanistically understand the abiotic and biotic controls on archaeal ecology.

The abiotic effects on archaeal community structure have been difficult to compare among studies, due partially to the reasons outlined above. Often temperature and pH are implicated as the predominant community structuring parameters, but the relative effect strength of the two is generally unclear (Bowen De Leon *et al.* 2013, Inskeep *et al.* 2005, Kozubal *et al.* 2012, Macur *et al.* 2004, Meyer-Dombard *et al.* 2011, Miller-Coleman *et al.* 2012, Swingley *et al.* 2012, Wang *et al.* 2013). Geochemical correlates to community structure have also been documented (Inskeep *et al.* 2005, Macur *et al.* 2004, Meyer-Dombard *et al.* 2005, Swingley *et al.* 2012), although sampling across large temperature or pH gradients may also contribute to or explain differences in community structure. Still, others have documented biogeographic or spring residence time as correlates to community structure (Bowen De Leon *et al.* 2013, Costa *et al.* 2009).

Two recent surveys across physicochemically heterogenous YNP springs including a metagenomic analysis of 20 springs (Inskeep *et al.* 2013b) and archaeal-only 16S rRNA gene analysis of 27 springs suggested that pH predominantly controls community composition, which reinforces earlier results implicating pH as one of the strongest correlates to hydrogenase distribution across YNP (Boyd *et al.* 2010) and bacterial community distribution across YNP (Mitchell 2009). Based on the phylogenetic and functional diversity in the aforementioned metagenomic data, Inskeep *et al.* also proposed a hierarchical model for environmental controls on thermal spring communities with pH first, followed by temperature, and various geochemical parameters (e.g. $[S^{2-}]$ and flow rate as a proxy for $[O_2]$). A recent metaanalysis of global thermal spring 16S rRNA gene data is also consistent with the general model proposed by Inskeep *et al.* (Xie *et al.* 2014).

In contrast, there is a much richer understanding of bacterial ecology in thermal springs, although many of the above issues regarding replication across relevant physicochemical parameters has often also been an issue. There, however, has not been large multi-domain censuses of both bacterial and archaeal phylogenetic diversity across relevant physicochemical gradients within YNP to date. Here we surveyed archaeal and bacterial communities among 32 physicochemically diverse YNP springs with the goals of 1) documenting archaeal phylogenetic diversity present in the YNP thermal spring ecosystem 2) elucidating the major abiotic drivers of archaeal community composition, 3) comparing diversity and distribution between Archaea and Bacteria and 4) investigating potential system-wide biotic interactions.

Materials and Methods

Sampling and Geochemical Analyses

Springs (n=32) were sampled in July of 2010 to include the range of temperature and pH of YNP thermal springs. Temperature and pH were measured using a Thermo Orion 290A+ meter (Thermo Fisher Scientific, Waltham, MA, USA) and conductivity was measured with a WTW meter (WTW, Wilheim, Germany) in the field. At each location, the planktonic community (~ 5-10 cm below the air-water interface) was sampled by filtering 0.6 - 1 L of spring water through a 0.22 uM pore size filter (Millipore, Bedford, MA, USA). Sediment communities, where present, were sampled at the same locale by sampling the top ~ 5 cm of sediment. Biomass was also sampled where present. Samples were preserved with sucrose lysis buffer (Giovannoni *et al.* 1990), which has been shown to preserve microbial biomass at ambient temperatures up to five days (Mitchell and Takacs-Vesbach 2008), then stored on ice in the field (< 2 days from collection) and stored in a freezer upon return to the laboratory. Dissolved S^{2-} was measured in triplicate in the field using the methylene blue method with a Hach DR 890 hand-held colorimeter (Hach Company, Loveland, CO, USA). Dissolved oxygen was measured in triplicate in the field using unfiltered water with Hach AccuVac evacuated ampules containing dissolved oxygen measurement reagents (Hach Company, Loveland, CO, USA). Filtered water was sampled for geochemical analysis and stored in deionized water washed bottles (for anion analysis) and acid-washed bottles (for cation analysis; also acidified with concentrated HNO_3^-). Major anions were measured using ion chromatography on a Dionex 500X Ion Chromatograph. Major cations and trace elements were measured using Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES; Perkin-Elmer, Waltham, MA, USA). Alkalinity was measured with titrations using

standardized sulfuric acid and calculated with the endpoint titration method (Pearson 1981). NH_4^+ was analyzed using the phenylhypochlorite-indophenol blue method (Koroleff 1983). Geochemical correlates were tested for autocorrelation to each other using pairwise Pearson product-moment correlations. Values below the detection limits of the procedure or instrument were given the minimum detectable limit for each analyte in correlational analyses.

DNA Extraction and Sequencing

DNA was extracted from all three sample types individually using a phenol/chloroform extraction method (Mitchell and Takacs-Vesbach 2008), modified to include a bead beating step. For each site, where present, ~0.5 g of sediment, ~200 uL of biomass (mat/filament), and the entire content of the planktonic biomass filter were extracted. Total community DNA for each community type was then pooled in equal volumes for each spring to represent the whole community of the spring sample site. While this approach necessarily combines potentially disparate communities with differing micro-scale geochemical attributes, the focus of the study was to document total spring community diversity. Additionally, because correlational analyses of geochemical analytes across habitat types (e.g. mat/sediment) among springs would incorporate communities not necessarily interacting with the same geochemical environment present in water used for analyses, a representation of the whole habitat-wide diversity of the community was chosen. Bacterial and archaeal 16S rRNA genes were sequenced using barcoded amplicon 454 pyrosequencing on a Roche 454 FLX instrument with Roche titanium reagents at Research and Testing Laboratory, Lubbock, TX. The bTEFAP protocol for 454 pyrosequencing has been described previously (Andreotti *et al.* 2011,

Dowd *et al.* 2008). Briefly, ~100 ng of DNA was amplified per sample in triplicate using universal primers with Roche-specific sequencing adapters (454 Life Sciences, Branford, CT, USA) and a barcode unique to each sample. The V3-V5 hypervariable regions were amplified for both Bacteria and Archaea using the bacterial-specific primers 341F (5' -CTACGGGAGGCAGCAG- 3') and 907R (5' -CCCCGTCAATTCCTTTGAGTT- 3') and archaeal-specific primers ARCH349F (5' -GYGCASCAGKCGMGAAW- 3') and ARCH806R (5' -GGACTACVSGGGTATCTAAT- 3') (Colman *et al.* 2015, Takai and Horikoshi 2000). Bacterial V3 amplifications were subject to a two-step amplification procedure, first amplifying near full-length 16S rRNA genes using universal primers 8F-1391R (Lane 1991) and attaching the barcodes to the amplicons in a second step. A subset of samples (n=15) were also amplified using bacterial-specific primers encompassing the V1-V3 hypervariable regions, 27F (5' -GAGTTTGATCNTGGCTCAG- 3') and 519R (5' -ATTACCGCGGCTGCTGG- 3') to compare diversity and taxonomic classifications between the two hypervariable regions.

DNA Sequence and Statistical Analyses

16S rRNA gene sequences were quality filtered, trimmed, denoised and screened for PCR errors and chimeras using AmpliconNoise and Perseus (Quince *et al.* 2011). Operational Taxonomic Units (OTUs) were then picked for each of the three datasets at the 97% similarity level, representative sequences were picked for each OTU and taxonomic affiliation was assigned for each OTU using the Quantitative Insights into Microbial Ecology (QIIME) pipeline (Caporaso *et al.* 2010). Sequences classified as Bacteria were filtered from the archaeal dataset, and sequences classified as Chloroplast or Archaea were filtered from the bacterial dataset. OTU tables were subsampled to 600

16S rRNA gene reads per sample in order to provide equitable sampling depths for comparisons.

Alpha diversity measurements were also performed in QIIME. OTU tables subsampled to 600 sequences were used to calculate alpha diversity measurements (Chao1 richness estimation and observed OTU richness) and Good's estimation of coverage. Statistical correlations of diversity measurements were conducted in the R environment (R Core Team 2014) on log transformed Chao1 estimates of richness, which provided normally distributed richness estimates. Total dataset rarefaction curves were calculated for each of the three 16S rRNA gene datasets using 600 subsampled reads from each spring. Beta diversity measurements were calculated in R with the vegan package (Oksanen *et al.* 2013) using Bray-Curtis distances to calculate community distances for archaeal and combined (V3 + V1) bacterial OTU tables. A Mantel test of V3 only and V1+V3 merged distance matrices suggested identical results (Mantel $r = 0.96$, $P \leq 0.001$), showing that the use of the additional V1 data did not influence beta-diversity results, but allowed greater inference of taxonomic classification abundances, which differed between V1 and V3. Non-metric multidimensional scaling (NMDS) ordinations were conducted using the metaMDS function in vegan and geochemical correlates were fitted to NMDS ordinations using the envfit function. Community dendrograms were constructed using an Unweighted Pair Group Method and Arithmetic mean (UPGMA) algorithm in hierarchical clustering. The co-occurrence of OTUs among the dataset were performed on a combined archaeal-bacterial non-singleton OTU table using the cooccur package for R (Griffith *et al.* 2014) which incorporates a probabilistic model for determining the statistical likelihood of species co-occurrences (Veech 2013).

Of the total significant ($P \geq 0.05$) co-occurrences that were found, results were further filtered to remove those that were between OTUs with the same taxonomic association and OTUs whose abundances across the dataset were not also significantly ($P \geq 0.05$) correlated by Spearman's r. Co-occurrence networks and network statistics were analyzed in Cytoscape (Smoot *et al.* 2011).

Results

Sampling, Geochemical Analysis and 16S rRNA Gene Sequencing Results

Thirty-two thermal springs were sampled in 14 thermal spring groups in nine geyser basins of Yellowstone National Park (Figure 3.1; Table 3.1). The temperature of the samples ranged from 31.2 to 88.4^{o}C and the pH ranged from 2.14 to 8.85. Most analytes, including the major ions, S^{2-}, NH_4^+, dissolved oxygen, As, and some metals varied over several orders of magnitude (Appendix File 3.A1). Many analytes were significantly correlated with pH and to a lesser extent; other parameters were significantly correlated with temperature (Figure 3.S1). Archaeal 16S rRNA gene data were recovered from all 32 samples, and bacterial 16S rRNA genes from all but one acidic site (Table 3.1). A total of 122,663 non-chimeric archaeal 16S rRNA gene reads (mean / sample = 3,717), 64,143 bacterial V3 reads (mean / sample = 2,138) and 85,091 bacterial V1 reads (mean / sample = 5,672) were produced.

Alpha Diversity and Taxonomic Composition Analyses

OTU diversity varied considerably among samples for all datasets (Table 3.2). Good's coverage estimation of diversity indicated that diversity was adequately sampled after subsampling each sample to 600 16S rRNA gene reads (range: 82 - 100% coverage, means: 96 - 97% coverage for all three datasets). Bacterial OTU diversity estimates based

on differing regions of the 16S rRNA gene were nearly identical, although some variation did exist between the two estimates (V3 vs. V1 log Chao1, $r^2 = 0.73$, n=14, β=0.98 by linear regression; $P \leq 0.01$). The diversity of Archaea was not significantly correlated with any of the measured environmental parameters. In contrast, bacterial diversity estimates were weakly, but significantly ($P \leq 0.05$) inversely correlated to temperature (Pearson's r = -0.53) and K (Pearson's r = -0.35).

Rarefaction curve analysis indicated a larger total system-wide diversity of Bacteria relative to Archaea based on both observed OTU richness and Chao1 richness estimates (Figure 3.2). The Chao1 estimated OTU richness difference between the two domains was smaller relative to the observed OTU richness difference. Both domains' rarefaction curves nearly reached asymptote (terminal rarefaction curve slopes = 0.023 and 0.019 for bacterial and archaeal observed OTU richness, respectively) indicating that there would be minimal increases in detected OTU richness with increased sampling. To examine evidence of niche partitioning of diversity between the two domains, the ratio of archaeal:bacterial diversity was tested against environmental parameters. Most springs contained similar levels of OTU richness, which was noted by normal distributions with means ~ 0 regardless of richness metric (Shapiro-Wilk normality test $P \geq 0.05$ for Chao1 and observed OTU richness; Figure 3.3). The distribution mean for the log Chao1 estimated diversity ratio was not significantly different from zero (indicating equal diversity between both domains among springs) by t test ($P > 0.05$, mean = -0.02, t=-0.34), but the observed OTU distribution mean was skewed towards higher bacterial diversity ($P = 0.05$, mean = -0.13, t=-2.014). The archaeal:bacterial diversity ratio was only significantly correlated to one environmental parameter: SO_4^{2-} (Pearson's r = 0.46).

Several other parameters were initially statistically significantly correlated to the diversity ratio (NH_4^+, Ca and Mg), but statistical significance was not observed after exclusion of the exceptionally archaeally diverse and high-leverage sample, SMH039 (which also harbored unique geochemistry).

The taxonomic composition for both domains varied considerably over the entire dataset. All four recognized archaeal phyla in the taxonomic reference set were present in the springs: Crenarchaeota (75.1%), Euryarchaeota (8.6%), Parvarchaeota (1.2%) and Nanoarchaeota (< 0.1%), and 15.1% archaeal reads that could not be assigned to a recognized phylum. Archaeal class definitions were used since widely accepted phyla such as the Thaumarchaeota and Korarchaeota are considered subphylum designations in the reference set that was used. Of the 27 classes in the reference set, 17 were found in the dataset, including 15.2% unclassified organisms, with the remaining 10 undetected groups primarily being Euryarchaeota and/or marine environment associated groups: Ancient Archaeal Group, Marine Benthic Group B, Marine Hydrothermal Vent Group, marine vent group pOWA133, Terrestrial Hot Spring Crenarchaeotic Group, Anaerobic Methanotrophic group, Halobacteria, Methanococci, Methanopyri and Thermococci. Thermoprotei constituted the largest overall percentage (47.7%), followed by the Aigarchaeota (15.2%), unclassified Archaea (15.2%), Thermoplasmata (8.2%) and the Thaumarchaeota (5.1%). The remaining groups each comprised < 4.0 % total abundance. The bacterial dataset consisted of 40 of the 85 phyla in the Greengenes reference set and a small percentage of unclassified reads (4.9% total). Thermi were the most abundant bacterial phylum (20.5%) followed by Aquificae (17.0%), Proteobacteria (16.0%) and Chloroflexi (9.2%). The rest of the phyla each constituted < 7.0% of the total.

Taxonomic Turnover and Beta Diversity Analyses

Turnover of taxonomic groups across the dataset was coincident with differences in pH and temperature. Archaeal community differences were organized primarily by pH and secondarily by temperature within similar pH profile groups (Figure 3.4). The highest temperature, most acidic springs, were dominated by Sulfolobales and Thermoproteales. Mid-temperature, high-acidity springs were comprised of Desulfurococcales, Thermoplasmatales, and the candidate divisions pUWA2 and pISA9. The lowest temperature acidic springs clustered with low-temperature slightly acidic springs and a single alkaline, low-temperature spring, MID059. The taxonomic distribution among these low temperature samples varied, with uncharacterized phyla dominating many of the springs along with Cenarchaeales in the highest pH spring and the candidate divisions YNPFFA and MBG in the acidic springs. The highest temperature mid- and high-pH springs clustered into two groups, which were differentiated by temperature. In the mid-temperature group, of which many were sampled from visible phototrophic growth, Aigarchaeota, uncharacterized phyla and Thaumarchaeota were the predominant taxonomic groups. At the highest temperature alkaline springs, Thermoproteales and Aigarchaeota constituted the majority of archaeal taxonomic abundance. A single Aigarchaeota OTU constituted 89 - 97% of all Aigarchaeota-classified reads in the highest temperature alkaline springs that had abundant Aigarchaeota populations (> 25% of the total community). However, that high-temperature Aigarchaeota OTU was not a large percentage (0 - 11%) of Aigarchaeota populations in mid-temperature alkaline springs where the Aigarchaeota were a major contributor to total community composition (>25% of the total community).

Bacterial communities were similarly structured by temperature and pH. The most acidic springs largely clustered into a single group that gradated by spring temperature (Figure 3.S2). In the highest temperature acidic springs, Proteobacteria, Aquificae and Firmicutes were the predominant phyla. In the lower temperature acidic springs, taxonomic distribution varied, with uncharacterized phyla, Acidobacteria, Proteobacteria and TPD-58 all constituting a large percentage in individual springs. Chloroflexi, Proteobacteria, Aquificae, Armatimonadetes and Chlorobi comprised most of the slightly acidic and alkaline mid-temperature springs, whereas Aquificae, Thermi and the candidate division EM3 dominated the higher temperature high pH springs.

Statistical analysis of the OTU distribution among sites confirmed that pH primarily structured communities, followed by temperature, though correlations to other geochemical variables were present. Non-metric multidimensional scaling (NMDS) ordinations of archaeal community dissimilarities structured samples into two groups along the first axis with the low pH sites as a single group and the mid- and high-pH sites as a second group. The sites were then structured along the second axis by temperature (Figure 3.5). The result was robust for both the archaeal and bacterial datasets. Both ordinations were supported by low stress values (0.13 and 0.16 for Archaea and Bacteria, respectively). A mantel test between archaeal and bacterial community distance matrices revealed a statistically significant and strong correlation for pairwise-site dissimilarities between the two domains (Mantel r=0.60, $P \leq 0.001$) The highest significantly correlated ($P \leq 0.05$) environmental factor to the samples in ordination space was pH ($r^2 = 0.87$ and 0.70 for Archaea and Bacteria, respectively) followed by a suite of geochemical parameters that were themselves highly correlated to pH, including total Fe, SO_4^{2-}, HCO_3^-,

Al, Ba, F, Zn, Na, As, B and others that were significantly correlated to temperature: Na, conductivity, K, Si, B, Br^- and Cl^- (Table 3.3). Temperature was also significantly correlated to both ordinations and the correlation was slightly stronger for Archaea (r^2 = 0.54, $P \leq 0.001$) than for Bacteria ($r^2 = 0.48$, $P < 0.001$). The only significant environmental correlate to community composition that was not also correlated to pH or temperature was NO_3^- ($r^2 = 0.26$ and 0.35 for Archaea and Bacteria, respectively).

OTU Co-occurrence Analyses

To investigate indicators of biotic interactions, statistically significant patterns of co-occurrence were measured. Of the 477,753 possible OTU combinations, inclusive of both domains, 540 statistically significant ($P \leq 0.05$) positive co-occurrences that were also significantly ($P \leq 0.05$) correlated by Spearman's r were found. Negative co-occurrence results were discarded, as biotic exclusion would be difficult to assess along the large abiotic gradients measured here. To filter the results, only co-occurrences between OTUs with different taxonomic classifications were used, as OTUs with identical taxonomic affiliation are most likely to be functionally equivalent. Slightly less than half of the co-occurrences were between members of different domains (42.4% of the 540 total). pH primarily structured OTU co-occurrences, with mean pH of OTU occurrence segregating samples into low and high pH sub networks (Figure 3.6a). The high/mid-pH sub network contained most of the OTU co-occurrence instances, and was investigated in greater detail (Figure 3.6b-d).

The phyla with the highest overall 16S rRNA gene abundances in the dataset contained OTUs with the highest degree (total number of significant co-occurrences) and the highest levels of closeness centrality (CC; proximity to all other nodes in the network;

Figure 3.6b). The five OTUs with the highest CC belonged to diverse phyla and comprised 19.7% of the total network edges, despite comprising 3.7% of the network's nodes (Table 3.4). Of the five nodes with the highest CC values, the only high-CC archaeal OTU was closely related (97% 16S rRNA gene sequence identity) to the Aigarchaeon, *Ca.* Caldiarchaeum subterraneum (Genbank accession# AP011878). The Caldiarchaeum-like organism, was however the OTU with the highest CC and node degree of both domains and was central to the high/mid-pH subnetwork (Figure 3.6c-d). A subnetwork only visualizing the nodes that were associated with Aigarchaeota provided insight into the Aigarchaeota-connected consortia. The Aigarchaeon OTU with the highest CC had significant associations with several OTUs from diverse phyla that were only associated with that OTU. The second highest CC for an OTU was for a Chloroflexi organism with high identity to the recently described *Thermoflexus hugenholtzii* (100% 16S rRNA gene identity to NR_125668; Dodsworth *et al.* 2014), which also shared OTU co-occurrences exclusively associated with it and the high CC Aigarchaeon. The following two highest CC OTUs shared high identity to *Thermus aquaticus* strain YT-1 (99% ID to NR_025900) and *Thermocrinis* sp. P2L2B (99% ID to AJ320219) and the fifth highest CC OTU was classified as an Armatimonadetes (formerly OP10) organism with low relationship to characterized organisms (85% ID to *Thermovenabulum ferriorganovorum* strain Z-9801).

Discussion

Alpha Diversity and Taxonomic Composition Analyses

Our results show that the wide variety of geochemical environments present in Yellowstone National Park (YNP) thermal springs supports a phylogenetically diverse

archaeal assemblage in diverse ecological habitats and that their communities are structured by environmental parameters, similarly, but also differentially than that of Bacteria. Most major archaeal classes were detected (except euryarchaeal and marine-associated lineages) in addition to uncharacterized lineages. The impetus to increase available genomic resources for uncharacterized and understudied lineages of Archaea and Bacteria has been previously noted (Rinke *et al.* 2013, Spang *et al.* 2015, Woyke and Rubin 2014, Wu *et al.* 2009) and our results reinforce that the YNP thermal spring ecosystem is an excellent resource for probing prokaryotic, and particularly, archaeal, diversity. In addition, our results expand a baseline distribution inventory for archaeal lineages throughout the YNP thermal spring ecosystem.

Overall archaeal richness was less than that of the Bacteria on a system-wide basis, but archaeal diversity was wider than expected; being present in every spring tested and being as rich, or richer than Bacteria in many springs (in all temperature and pH profiles). This is consistent with previous estimates that Archaea are likely to be most diverse in hydrothermal systems (Auguet *et al.* 2010, Barns *et al.* 1994). Importantly, archaeal richness wasn't partitioned only into the most 'extreme' sites (e.g. in terms of temperature or pH), as would fit the traditional perspective of their relative dominance in the most marginal habitats (Reysenbach and Shock 2002, Robertson *et al.* 2005). It should, however, be noted that 16S rRNA gene diversity does not reflect *in situ* population abundances, and thus although most springs harbored similar levels of archaeal and bacterial diversity, community abundances may still be skewed towards bacterial dominance. Archaea still likely constitute subdominant populations to Bacteria in most sites where they are as much, or more diverse than the cohabitating Bacteria.

Previous studies (Hugenholtz *et al.* 1998, Spear *et al.* 2005) have suggested Archaea are subdominant to Bacteria in thermal springs, although acidic springs may be the exception (Inskeep *et al.* 2013b). The only other deep sequencing archaeal survey effort across large physicochemical gradients in YNP to report richness estimates was recently conducted across 27 YNP thermal springs (Boyd *et al.* 2013). While Boyd *et al.* found an overall much lower level of archaeal richness than we found here (33 total OTUs from 22 samples), our dataset comprises an order of magnitude more 16S rRNA gene reads and increased site sampling. An analysis of 16 Yunnan/Tibet thermal springs revealed higher OTU diversity than was found here, despite having overall smaller sampling effort, but the quality filtering and data subsampling methods differed (Song *et al.* 2013). Nevertheless, sampling protocol, sequencing read depth, sequencing methodology and quality-filtering steps among other factors can influence absolute values of richness, which makes accurate comparisons of richness across studies difficult.

The archaeal taxonomic composition among the springs was largely consistent with expected ecological distributions based on the physiological properties of characterized taxonomic lineages, although several uncharacterized or newly characterized lineages provided opportunities to understand their ecology. The dominance of the Thermoprotei class across the dataset was consistent with their dominance of YNP springs reported in previous studies (Inskeep *et al.* 2013a, Meyer-Dombard *et al.* 2005, Spear *et al.* 2005) and Yunnan/Tibetan thermal springs (Song *et al.* 2013). Within the order Thermoprotei, the partitioning of characterized orders into distinct pH groups was consistent with their physiology. The abundance-weighted temperature and pH means for Sulfolobales, Thermoproteales and Desulfurococcales

presence was consistent with their physiological tolerances as acidophiles, weak acidophiles/neutrophiles, and pH-variable lineages, respectively, in extremely thermophilic environments (Huber *et al.* 2006, Huber and Prangishvaili 2006, Huber and Stetter 2006). The YNPFFA Thermoprotei clade was only found in significant abundance (> 5.0% of community) in three sites that were acidic (pH 2.1 - 4.0) and low temperature (42.3 - 66.0 ^{o}C). The YNPFFA are thus far only known from 16S rRNA gene surveys, but this result is consistent with their phylogenetic association with other mesophilic Thaumarchaeota (Brochier-Armanet *et al.* 2008, Schleper *et al.* 2005).

Distribution results for recently characterized lineages provided insight into their ecological distributions. The ubiquity of the proposed Aigarchaeota (formerly the pSL4 candidate division) among high temperature neutral-alkaline sites has not been documented for YNP springs, although that may be in part due to their classification as 'Unclassified Crenarchaea' in previous studies (e.g. Meyer-Dombard *et al.* 2011). They have been found in high abundance in some high temperature and high pH springs in some Great Basin (Nevada, US) and Heart Lake Geyser Basin springs (YNP, US) (Bowen De Leon *et al.* 2013, Cole *et al.* 2013). Recently characterized Aigarchaeota genomes (Hedlund *et al.* 2014, Nunoura *et al.* 2011) from uncultured representatives indicate that they possess the genomic potential for autotrophy through the dicarboxylate/4-hydroxybutyrate cycle. Given the widespread distribution of the Aigarchaeota among neutral-alkaline high-temperature sites here, they may fill a previously unappreciated role in the productivity of high temperature, circumneutral thermal spring communities. The segregation of different Aigarchaeota OTUs into different temperature profiles in alkaline springs also suggests a potential diversity of

Aigarchaeota ecotypes within the YNP thermal springs system. The role of ammonia-oxidizing Thaumarchaea in thermal spring nutrient cycling (and global nutrient cycling) has also been given considerable attention since their characterization (Brochier-Armanet *et al.* 2008, de la Torre *et al.* 2008, Dodsworth *et al.* 2011). The Cenarchaeales were only found in one spring in any significant abundance (>5.0%) and the mesophilic profile of the spring (T=39^{o}C, pH=8.7) is consistent with the dominance of Cenarchaeales (particularly *Nitrosopumilus* spp.) in mesophilic aquatic environments (Karner *et al.* 2001, Konneke *et al.* 2005). The Nitrosocaldales (inclusive of the only cultivated member, *Ca.* Nitrosocaldus yellowstonii; de la Torre *et al.* 2008) were only found in alkaline springs with moderately thermophilic temperatures. Unlike a recent survey of YNP springs (Boyd *et al.* 2013), the Nitrosocaldales were not a large percentage of our overall taxonomic diversity (3.1% of total), but were restricted in high abundance (> 5.0%) to sites with visible photosynthetic mats (sans LOW021). It's possible that the Nitrosocaldales abundance in sites with mats is attributable to being present primarily in alkaline-chloride sites with typically low NH_4^+ that may favor Thaumarchaea that have high affinities for NH_4^+ (Hamilton *et al.* 2014) and also fall within the optimal growth temperature range as indicated by the only Nitrosocaldales cultivated member (de la Torre *et al.* 2008). An alternative hypothesis is that the Thaumarchaea contribute to community assembly through competition for bioavailable nitrogen (NH_4^+), preferentially recruiting diazotrophs into the community, as has been proposed (Hamilton *et al.* 2014). However, of the dominant bacterial OTUs in the five sites with appreciable Thaumarchaeal abundance (*Roseiflexus* spp., *Chloroflexus* spp. and *Thermus* spp.), none are known diazotrophs. It's possible that this biotic interaction may be occurring with

subdominant community members, such as the *Thermocrinis*-affiliated organisms in LOW021; a genus of which some species contain the gene repertoire for diazotrophy (Boyd and Peters 2013).

Biodiversity surveys of relatively understudied groups, such as the Archaea, provide excellent baseline data from which to target phylogenetic lineages of ecological or evolutionary importance for genomic analyses (Rinke *et al.* 2013). Most of the archaeal 16S rRNA gene reads that could not be assigned to phyla (15% of all archaeal reads) were found in moderately thermophilic sites with circumneutral pH (Figure 3.4). Many of these sites also harbored visible photosynthetic mats. One explanation for this could be the historical bias of archaeal studies to geothermal habitats that are restricted to the highest temperatures or acidic springs (where they tend to be most abundant). Even likely being at low abundances, they appear to have phylogenetically diverse communities in circumneutral springs, despite that they are often thought to be entirely absent or minor members of communities in these sites where Bacteria are visibly dominant (Inskeep *et al.* 2010, Miller *et al.* 2009, Ward *et al.* 1998). A recent analysis of three alkaline springs in the Heart Lake Geyser Basin of YNP supports that there is a relatively unappreciated diversity and ubiquity of Archaea in moderately thermophilic circumneutral springs (Bowen De Leon *et al.* 2013). When considering taxonomic ranks below the phylum/class level, the number of 16S rRNA gene sequences not classifiable into known orders (23.5%) highlights the considerable archaeal phylogenetic diversity that is uncharacterized in the YNP thermal spring system.

The Korarchaeota were only present in any significant abundance (>1.0%) at one site (SMH039), which was surprising given their consistent presence in terrestrial thermal

springs, including YNP (Auchtung *et al.* 2006, Auchtung *et al.* 2011, Miller-Coleman *et al.* 2012, Reigstad *et al.* 2010). A recent metagenomic analysis of 20 YNP thermal springs also failed to find any appreciable Korarchaeal populations in all but a single site (Inskeep *et al.* 2013a). This result, combined with the data here (and other estimates indicating low population abundances in samples where they are found: Auchtung *et al.* 2011, Reigstad *et al.* 2010) supports a discrete and marginal inter-spring distribution for the Korarchaeota among thermal fields. Conversely, Nanoarchaeota were only found in very minor abundances (< 0.1%) in three samples, despite that they may be widespread in YNP springs (Inskeep *et al.* 2013a). This lower 16S rRNA gene abundance may indicate preferential amplification bias against the Nanoarchaeota, as has been indicated for most archaeal-specific PCR primer pairs (Klindworth *et al.* 2013).

Taxonomic Turnover and Beta Diversity Analyses

pH and temperature were the strongest statistical correlates to bacterial and archaeal community composition. Spring pH correlated to a discrete difference in community composition (springs below pH 5 for both Archaea and Bacteria largely separating from springs with pH > 5) and was by far the strongest correlate to differences in community composition. Within discrete pH categories (low/high), temperature further separated communities. The strong influence of pH on community composition has been noted for Bacteria, Archaea, functional gene diversity and metagenomic communities in thermal springs (Boyd *et al.* 2010, Inskeep *et al.* 2013b, Mitchell 2009, Sharp *et al.* 2014, Xie *et al.* 2014) and in other mesophilic environments (Lauber *et al.* 2009, Lindstrom *et al.* 2005). While this discrete separation could be reflective of the bimodal distribution of pH in YNP springs with relatively few springs in the pH 4-5 range and two distribution

peaks near ~2.5 and 6.5 (Inskeep *et al.* 2013b), it is also likely reflective of the physiological tolerances of the guilds that comprise the communities, given the wide range of pH sampled here. As noted above for Thermoprotei, many taxonomic groups were distributed along known physiological tolerances.

There were also statistically significant geochemical correlates to community composition. However, it is difficult to statistically assess the influence of geochemistry on community composition when the large scale of pH and temperature on communities masks the subtler signal of geochemical correlates. The issue is exacerbated because pH and temperature controls geochemical characteristics (e.g. increased metal content in acidic springs from leaching), is reflective of geological processes that in turn effect spring chemistry (e.g. vapor-dominated systems where gases can contribute to acidity through oxidation) and affects the energetics of potential metabolic reactions (Nordstrom *et al.* 2005, Shock *et al.* 2010). Another consideration is that the whole spring community approach used here may aggregate functional and taxonomic differences that can exist across distinct habitat phases in thermal springs (Murphy *et al.* 2013), where for instance, oxygen availability may differ on the mm scale even within habitat types like photosynthetic mats (Ward *et al.* 2006). Consequently, geochemical data taken at the sampling locale from water, as done here, may only reflect the strongest geochemical differences among springs, and an accurate statistical approach to inferring the role of geochemistry in community structure is needed by deep sampling within narrow temperature and pH profiles that should also consider the heterogeneity of chemical gradients at the mm scale.

Qualitative differences of communities within the same physicochemical profile can, however, provide insight into the role of geochemistry and geologic processes on archaeal communities. For example, despite that LCB011 and SMH039 shared nearly identical temperature (68.8, 68.2) and pH profiles (6.95, 7.06), both shared highly dissimilar archaeal and bacterial communities coincident with large geochemical differences. The $>10^2$ higher S^{2-} and NH_4^+ in SMH039 than LCB011 (both below detection in LCB011) reflects the differing geologic processes that contribute to their spring chemistry. SMH039 is in the Seven Mile Hole region of northeastern YNP which has abundant vapor-influenced systems with typically high NH_4^+ and S^{2-}, while LCB011 is located in the Lower Geyser Basin and was typical of alkaline-chloride springs that are ubiquitous in this region (e.g. Na/Cl dominated, low S^{2-}, circumneutral and high Si; Fournier 1989). LCB011 (a sample with visible photosynthetic mat) had less genera level-classifications of Archaea than SMH039 and primarily contained *Ca.* Caldiarchaeum spp. and *Ca.* Nitrosocaldus spp. while SMH039 (a sample with visible yellow filaments along the spring runoff channel) contained primarily unclassified Archaea, *Pyrobaculum* spp., and unclassified organisms within the Parvarchaeota and Aigarchaeota divisions (Figure 3.S3). Bacteria were also highly dissimilar between sites with LCB011 being more even and primarily being composed of *Roseiflexus* spp. and an unclassified genus within the Armatimonadetes, while SMH039 had primarily Hydrogenothermaceae-affiliated organisms and *Thermus* spp. The exclusion of photosynthetic bacteria in high S^{2-} springs that are otherwise appropriate habitats by temperature (below the observational upper temperature limit for photosynthesis of ~74^{o}C) and pH has been noted previously in YNP springs (Cox *et al.* 2011). Our results

suggest a similar community structuring effect on Archaea, although further studies are necessary to discern whether the mechanism underlying it are abioticially or bacterially mediated. Nonetheless, there are apparent correlations of geochemical processes on community composition, but the large effects of temperature and pH community composition necessitate sampling across geochemically diverse regions while considering temperature and pH differences to adequately statistically test the magnitude of the correlation.

Generally, both domains' communities were similarly structured in response to environmental parameters. However, there were some differences that may highlight potential niche partitioning between the two domains. Bacterial richness declined with temperature overall, which is consistent with a species filtering effect of temperature on bacterial lineages that others have reported in individual springs (Miller *et al.* 2009). The correlation of K to bacterial richness can also be explained by factors at least partially related to temperature. K was significantly correlated to temperature in our dataset, and therefore may itself not be correlated to bacterial diversity but rather, an indication of high temperature systems. Archaeal richness itself was not correlated to any parameters, but the ratio of archaeal richness:bacterial richness was significantly correlated to a single parameter (SO_4^{2-}) that hints at niche partitioning along physical parameters. The three sites with the highest archaeal:bacterial richness (APT001, SMH039, LOW023) span a large range of pH (2.2, 7.06 and 5.93, respectively), are high temperature systems (69.5, 68.2 and 73.5) and exhibit either high levels of oxidized (SO_4^{2-}) or reduced (S^{2-}) sulfur which suggests they are impacted by processes that contribute reduced sulfur to the system that is oxidized upon mixing with oxygenated near-surface waters (Fournier

1989). Because SO_4^{2-} was significantly correlated to pH, but pH itself wasn't correlated to the between-domain diversity ratio, the result indicates geochemical processes that underlie the contribution of large amounts of SO_4^{2-} to the system may favor the presence of more diverse archaeal assemblages relative to Bacteria. Further supporting this hypothesis is that we were unable to amplify Bacteria from a single site: an acid-sulfate-chloride spring (NOR006; Cinder Pool) with high levels of SO_4^{2-} (0.97 mM). Cinder Pool exhibits atypical sulfur geochemistry for acid-sulfate springs due to a shallow (~18 m) pool of molten sulfur below the spring surface (Xu *et al.* 2000).

OTU Co-occurrence Analyses

While abiotic parameters are clearly important in structuring microbial communities of thermal springs, biotic interactions that affect community composition have received less attention. Although photosynthetic mat consortia have been studied in detail (reviewed in: Ward *et al.* 1998), and single spring/runoff channel community interactions have also been studied (Hall *et al.* 2008, Hamilton *et al.* 2014, Miller *et al.* 2009); system-wide, across-domain interactions have recieved little attention. Our dataset permits some speculation into community-wide interactions across the YNP region by comparing statistically significant correlations of abundance and the co-occurrence of OTUs. Though competitive exclusion is likely also occurring, it's difficult to detect exclusion when large abiotic gradients may obscure negative interaction signals, and thus negatively correlated occurrences were ignored here. Likewise, correlated presence does not necessarily imply biotic interaction, but may also suggest overlapping niches across relevant physicochemical parameters. It should also be noted that there are inherent limitations to inferring network-based interactions from 16S rRNA gene abundance data,

which inherently bias abundance estimates through PCR primer taxonomic biases, unequal 16S rRNA gene copy numbers in taxa, and contributions from dormant populations among others (Lupatini *et al.* 2014). Nevertheless, identifying putative interactions can provide testable hypotheses with which to understand how thermal spring community composition is structured biotically and aid identification of important contributors to community structure and function across the YNP thermal spring system.

OTUs were largely only associated with others within similar pH profiles (Figure 3.6a), which further reinforced the strong filtering effect of pH on communities and the restriction of theoretical interactions to physicochemically comparable springs. A considerable fraction of the correlated occurrences were across domains and also reinforces the widespread sharing of habitats between Archaea and Bacteria and suggests similar guilds of both domains inhabit springs across the YNP thermal ecosystem. An Aigarchaeota-associated OTU had the highest closeness centrality (CC quantifies the distance from one node to all others in a network), which suggests that they could be a previously unappreciated and important component of circumneutral to alkaline pH spring communities in YNP. The only characterized Aigarchaeota genomes (Hedlund *et al.* 2014, Nunoura *et al.* 2011) indicate a metabolic potential for autotrophy, and thus they may contribute to the productivity of thermal spring consortia in YNP. The network connections between the Aigarchaeota OTU and a diverse array of Bacteria and Archaea (including several that are only correlated to the high CC Aigarchaeon OTU) support this hypothesis. It should be noted that of the other 5 highest CC OTUs, the only other likely autotroph was an OTU classified as a *Thermocrinis* sp. closely related to the autotrophic *T. ruber* (Huber 1998). Two of the other high CC nodes, the recently characterized

Chloroflexi *Thermoflexus hugenholtzii* and *Thermus aquaticus,* are both heterotrophic (Da Costa *et al.* 2006, Dodsworth *et al.* 2014), and the lack of close relatives to the Armatimonadetes-classified OTU prohibits functional inference. One set of hypotheses as to the high centrality of these OTUs to the network is that they either contribute to primary productivity (e.g. *Thermocrinis* spp. and possibly Aigarchaeota-related OTUs) or are capable of establishing ubiquitous, opportunistic heterotophic populations that are dependent on other organisms for nutrient availability (e.g. *T. aquaticus*. and *T. hugenholtzii*). Of the OTUs with the ten highest CC, most were related to characterized phyla, except for an archaeon that could not be assigned to a phylum, that was closely related to an environmental clone from the Perpetual Spouter spring of YNP (95% nt ID; KC254665; Hamilton *et al.* 2014), but $\leq$ 84% to other uncultured clones and $\leq$ 76% to characterized Euryarchaeotes. Intriguingly, the highest correlated OTU to the unidentified archaeon was a *Ca.* Nitrosocaldus-related organism (Spearman's $r = 0.87$, $P < 0.05$), which were also dominant community members in Perpetual Spouter spring. In Perpetual Spouter, *Ca.* Nitrosocaldus spp. may influence community composition through limiting available NH_4^+, which in turn may select for diazotrophic members of the community (Hamilton *et al.* 2014). The high correlation between the unclassified OTU and the *Ca.* Nitrosocaldus classified OTU suggests either biotic interaction among multiple springs, potentially through N availability, or simply that they inhabit similar physicochemical niches among the YNP thermal spring ecosystem. The high centrality of the unclassified OTU suggests that they may be ecologically important, entirely uncharacterized components of circumneutral thermal spring consortia and provide an

intriguing target for future analyses of biotic interaction studies and genomic characterization.

Conclusions

Here we report a previously unappreciated diversity of Archaea in YNP thermal springs at all temperature and pH profiles including most of the major lineages of Archaea. Although the overall diversity of Archaea in the YNP system was not as high as that of the Bacteria, this may be skewed by the significant enrichment of bacterial lineages (and the converse paucity of archaeal lineages) in many of the lowest temperature sites. In most of the springs sampled here, the Archaea were as much, or more diverse than Bacteria within the same springs. Both domains' communities are correlated to environmental parameters similarly, with pH, followed by temperature being the predominant correlates to community composition. There was qualitative evidence for geochemically-mediated community structuring effects on both domains. However, targeted sampling of geochemically heterogenous springs within narrow temperature and pH ranges is needed to accurately quantify the effect of geochemical factors on communities across the YNP system. There were myriad putative interactions among and within domains, and network analyses suggested potentially important contributions to community structure from uncharacterized or only recently characterized Archaea, including the Aigarchaeota. Future targeted analyses upon the baseline distribution and diversity survey results reported here may help untangle how biotic and abiotic parameters themselves interact to ultimately establish the diverse thermal spring archaeal and bacterial consortia detected here.

Acknowledgements

This work was supported by grants from the National Science Foundation (NSF DEB0206773) and the UNM Research Allocation Committee to CTV. DRC received additional support from NSF EHR0832947 and the UNM Office of Graduate Studies Research, Project and Travel Grant in addition to the UNM OGS Graduate Student Success Scholarship. We thank Beverly Marrs for field assistance, the Yellowstone Center for Resources (National Park Service), Christie Hendrix and Stacy Gunther for permitting and field assistance.

References

Andreotti R, Pérez de León AA, Dowd SE, Guerrero FD, Bendele KG, Scoles GA (2011). Assessment of bacterial diversity in the cattle tick *Rhipicephalus* (*Boophilus*) *microplus* through tag-encoded pyrosequencing. *BMC Microbiol* **11:** 6.

Auchtung TA, Takacs-Vesbach CD, Cavanaugh CM (2006). 16S rRNA phylogenetic investigation of the candidate division "Korarchaeota". *Appl Environ Microbiol* **72:** 5077-5082.

Auchtung TA, Shyndriayeva G, Cavanaugh CM (2011). 16S rRNA phylogenetic analysis and quantification of Korarchaeota indigenous to the hot springs of Kamchatka, Russia. *Extremophiles* **15:** 105-116.

Auguet JC, Barberan A, Casamayor EO (2010). Global ecological patterns in uncultured Archaea. *ISME J* **4:** 182-190.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci U S A* **107:** 8806-8811.

Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci U S A* **91:** 1609-1613.

Bowen De Leon K, Gerlach R, Peyton BM, Fields MW (2013). Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Front Microbiol* **4:** 330.

Boyd ES, Hamilton TL, Spear JR, Lavin M, Peters JW (2010). [FeFe]-hydrogenase in Yellowstone National Park: evidence for dispersal limitation and phylogenetic niche conservatism. *ISME J* **4:** 1485-1495.

Boyd ES, Hamilton TL, Wang J, He L, Zhang CL (2013). The role of tetraether lipid composition in the adaptation of thermophilic archaea to acidity. *Front Microbiol* **4:** 62.

Boyd ES, Peters JW (2013). New insights into the evolutionary history of biological nitrogen fixation. *Front Microbiol* **4:** 201.

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008). Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* **6:** 245-252.

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

Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ *et al* (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* **25:** 690-701.

Cole JK, Peacock JP, Dodsworth JA, Williams AJ, Thompson DB, Dong HL *et al* (2013). Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. *ISME J* **7:** 718-729.

Colman DR, Thomas R, Maas KR, Takacs-Vesbach CD (2015). Detection and analysis of elusive members of a novel and diverse archaeal community within a thermal spring streamer consortium. *Extremophiles* **19:** 307-315.

Costa KC, Navarro JB, Shock EL, Zhang CLL, Soukup D, Hedlund BP (2009). Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. *Extremophiles* **13:** 447-459.

Cox A, Shock EL, Havig JR (2011). The transition to microbial photosynthesis in hot spring ecosystems. *Chem Geol* **280:** 344-351.

Da Costa MS, Rainey FA, Nobre MF (2006). The Genus Thermus and Relatives. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

de la Torre JR, Walker CB, Ingalls AE, Konneke M, Stahl DA (2008). Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* **10:** 810-818.

Dodsworth JA, Hungate BA, Hedlund BP (2011). Ammonia oxidation, denitrification and dissimilatory nitrate reduction to ammonium in two US Great Basin hot springs with abundant ammonia-oxidizing archaea. *Environ Microbiol* **13:** 2371-2386.

Dodsworth JA, Gevorkian J, Despujos F, Cole JK, Murugapiran SK, Ming H *et al* (2014). Thermoflexus hugenholtzii gen. nov., sp. nov., a thermophilic, microaerophilic, filamentous bacterium representing a novel class in the Chloroflexi, Thermoflexia classis nov., and description of Thermoflexaceae fam. nov. and Thermoflexales ord. nov. *Int J Syst Evol Microbiol* **64:** 2119-2127.

Dowd SE, Wolcott RD, Sun Y, McKeehan T, Smith E, Rhoads D (2008). Polymicrobial nature of chronic diabetic foot ulcer biofilm infections determined using bacterial tag encoded FLX amplicon pyrosequencing (bTEFAP). *PLoS One* **3:** e3326.

Fournier RO (1989). Geochemistry and Dynamics of the Yellowstone-National-Park Hydrothermal System. *Annu Rev Earth Pl Sc* **17:** 13-53.

Giovannoni SJ, DeLong EF, Schmidt TM, Pace NR (1990). Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Appl Environ Microb* **56:** 2572-2575.

Griffith D, Veech J, Marsh C (2014). cooccur: Probabilistic Species Co-occurrence Analysis in R.

Hall JR, Mitchell KR, Jackson-Weaver O, Kooser AS, Cron BR, Crossey LJ *et al* (2008). Molecular Characterization of the Diversity and Distribution of a Thermal Spring Microbial Community by Using rRNA and Metabolic Genes. *Appl Environ Microbiol* **74**: 4910-4922.

Hamilton TL, Koonce E, Howells A, Havig JR, Jewell T, de la Torre JR *et al* (2014). Competition for ammonia influences the structure of chemotrophic communities in geothermal springs. *Appl Environ Microbiol* **80:** 653-661.

Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles* **18:** 865-875.

Huang Q, Dong CZ, Dong RM, Jiang H, Wang S, Wang G *et al* (2011). Archaeal and bacterial diversity in hot springs on the Tibetan Plateau, China. *Extremophiles* **15:** 549-563.

Huber H, Huber R, Stetter KO (2006). Thermoproteales. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

Huber H, Prangishvaili D (2006). Sulfolobales. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

Huber H, Stetter KO (2006). Desulfurococcales. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* **180:** 366-376.

Inskeep WP, Ackerman GG, Taylor WP, Kozubal M, Korf S, Macur RE (2005). On the energetics of chemolithotrophy in nonequilibirium systems: case studies of teothermal springs in Yellowstone National Park. *Geobiology* **3:** 297-317.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH *et al* (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5:** e9773.

Inskeep WP, Jay ZJ, Herrgard MJ, Kozubal MA, Rusch DB, Tringe SG *et al* (2013a). Phylogenetic and Functional Analysis of Metagenome Sequence from High-Temperature Archaeal Habitats Demonstrate Linkages between Metabolic Potential and Geochemistry. *Front Microbiol* **4:** 95.

Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB, YNP Metagenome Project Steering Committee *et al* (2013b). The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. *Front Microbiol* **4:** 67.

Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409:** 507-510.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M *et al* (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41:** e1.

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437:** 543-546.

Koroleff F (1983). Determination of Ammonia. In: Grasshoff K, Ehrhardt M, Kremling K (eds). *Methods of Seawater Analysis*, 2nd edn. Wiley-VCH: Weinheim. pp 150-157.

Kozubal MA, Macur RE, Jay ZJ, Beam JP, Malfatti SA, Tringe SG *et al* (2012). Microbial iron cycling in acidic geothermal springs of yellowstone national park: integrating molecular surveys, geochemical processes, and isolation of novel fe-active microorganisms. *Front Microbiol* **3:** 109.

Kozubal MA, Romine M, Jennings R, Jay ZJ, Tringe SG, Rusch DB *et al* (2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J* **7:** 622-634.

Lane DJ (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic acid techniques in bacterial systematics*. John Wiley & Sons: Chichester, United Kingdom. pp 115-175.

Lauber CL, Hamady M, Knight R, Fierer N (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75:** 5111-5120.

Lindstrom ES, Kamst-Van Agterveld MP, Zwart G (2005). Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. *Appl Environ Microbiol* **71:** 8201-8206.

Lupatini M, Suleiman AKA, Jacques RJS, Antoniolii ZI, de Siqueira Ferreira A, Kuramae EE *et al* (2014). Network topology reveals high connectance levels and few key microbial genera within soils. *Frontiers in Environmental Science* **2:** 1-11.

Macur RE, Langner HW, Kocar BD, Inskeep WP (2004). Linking geochemical processes with microbial community analysis: successional dynamics in an arsenc-rich, acid-sulphate-chloride geothermal spring. *Geobiology* **2:** 163-177.

Meyer-Dombard DR, Shock EL, Amend JP (2005). Archaeal and bacterial communities in geochemically diverse hot springs of Yellowstone National Park, USA. *Geobiology* **3:** 211-227.

Meyer-Dombard DR, Swingley W, Raymond J, Havig J, Shock EL, Summons RE (2011). Hydrothermal ecotones and streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. *Environ Microbiol* **13:** 2216-2231.

Miller SR, Strong AL, Jones KL, Ungerer MC (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. *Appl Environ Microbiol* **75:** 4565-4572.

Miller-Coleman RL, Dodsworth JA, Ross CA, Shock EL, Williams AJ, Hartnett HE *et al* (2012). Korarchaeota diversity, biogeography, and abundance in Yellowstone and Great Basin hot springs and ecological niche modeling based on machine learning. *PLoS One* **7:** e35964.

Mitchell KR, Takacs-Vesbach CD (2008). A comparison of methods for total community DNA preservation and extraction from various thermal environments. *J Ind Microbiol Biotechnol* **35:** 1139-1147.

Mitchell KR (2009). Controls on microbial community structure in thermal environments; exploring Bacterial diversity and the relative influence of geochemistry and geography. Ph.D. thesis, University of New Mexico, Albuquerque.

Murphy CN, Dodsworth JA, Babbitt AB, Hedlund BP (2013). Community microrespirometry and molecular analyses reveal a diverse energy economy in Great Boiling Spring and Sandy's Spring West in the U.S. Great Basin. *Appl Environ Microbiol* **79:** 3306-3310.

Nordstrom KD, Ball JW, McCleskey RB (2005). Ground water to surface water: chemistry of thermal outflows in Yellowstone National Park. *Geothermal biology and geochemistry in Yellowstone National Park***:** 73 - 94.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al* (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39:** 3204-3223.

Offre P, Spang A, Schleper C (2013). Archaea in Biogeochemical Cycles. *Annu Rev Microbiol, Vol 67* **67:** 437-457.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB *et al* (2013). vegan: Community Ecology Package, R package version 2.0-10 edn.

Pearson F (1981). Fixed Endpoint Alkalinity Determintation. *J Water Pollut Con F***:** 1243-1252.

Perevalova AA, Kolganova TV, Birkeland NK, Schleper C, Bonch-Osmolovskaya EA, Lebedinsky AV (2008). Distribution of Crenarchaeota representatives in terrestrial hot springs of Russia and Iceland. *Appl Environ Microbiol* **74:** 7620-7628.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing Noise From Pyrosequenced Amplicons. *Bmc Bioinformatics* **12**.

Reigstad LJ, Jorgensen SL, Schleper C (2010). Diversity and abundance of Korarchaeota in terrestrial hot springs of Iceland and Kamchatka. *ISME J* **4:** 346-356.

Reysenbach AL, Shock E (2002). Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296:** 1077-1082.

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

Robertson CE, Harris JK, Spear JR, Pace NR (2005). Phylogenetic diversity and ecology of environmental Archaea. *Curr Opin Microbiol* **8:** 638-642.

Schleper C, Jurgens G, Jonuscheit M (2005). Genomic studies of uncultivated archaea. *Nat Rev Microbiol* **3:** 479-488.

Sharp CE, Brady AL, Sharp GH, Grasby SE, Stott MB, Dunfield PF (2014). Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. *ISME J* **8:** 1166-1174.

Shock EL, Holland M, Meyer-Dombard D, Amend JP, Osburn GR, Fischer TP (2010). Quantifying inorganic sources of geochemical energy in hydrothermal ecosystems, Yellowstone National Park, USA. *Geochim Cosmochim Ac* **74:** 4005-4043.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27:** 431-432.

Song ZQ, Wang FP, Zhi XY, Chen JQ, Zhou EM, Liang F *et al* (2013). Bacterial and archaeal diversities in Yunnan and Tibetan hot springs, China. *Environ Microbiol* **15:** 1160-1175.

Spang A, Martijn J, Saw JH, Lind AE, Guy L, Ettema TJ (2013). Close encounters of the third domain: the emerging genomic view of archaeal diversity and evolution. *Archaea* **2013:** 202358.

Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE *et al* (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521:** 173-179.

Spear JR, Walker JJ, McCollom TM, Pace NR (2005). Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci U S A* **102:** 2555-2560.

Swingley WD, Meyer-Dombard DR, Shock EL, Alsop EB, Falenski HD, Havig JR *et al* (2012). Coordinating environmental genomics and geochemistry reveals metabolic transitions in a hot spring ecosystem. *PLoS One* **7:** e38108.

Takai K, Horikoshi K (1999). Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* **152:** 1285-1297.

Takai K, 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.

Team RC (2014). R: A language and environment for statistical computing, 3.1.0 edn. R Foundation for Statistical Computing: Vienna, Austria.

Veech JA (2013). A probabilistic model for analysing species co-occurrence. *Global Ecol Biogeogr* **22:** 252-260.

Wang S, Hou W, Dong H, Jiang H, Huang L, Wu G *et al* (2013). Control of temperature on microbial community structure in hot springs of the Tibetan Plateau. *PLoS One* **8:** e62901.

Ward DM, Ferris MJ, Nold SC, Bateson MM (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol Mol Biol Rev* **62:** 1353-1370.

Ward DM, Bateson MM, Ferris MJ, Kuhl M, Wieland A, Koeppel A *et al* (2006). Cyanobacterial ecotypes in the microbial mat community of Mushroom Spring (Yellowstone National Park, Wyoming) as species-like units linking microbial community composition, structure and function. *Philos T R Soc B.* **361:** 1997-2008.

Woese CR, Fox GE (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74:** 5088-5090.

Woyke T, Rubin EM (2014). Evolution. Searching for new branches on the tree of life. *Science* **346:** 698-699.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN *et al* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462:** 1056-1060.

Xie W, Zhang CL, Wang J, Chen Y, Zhu Y, de la Torre JR *et al* (2014). Distribution of ether lipids and composition of the archaeal community in terrestrial geothermal springs: impact of environmental variables. *Environ Microbiol* **17**: 1600-1614.

Xu Y, Schoonen MAA, Nordstrom DK, Cunnigham KM, Ball JW (2000). Sulfur geochemistry of hydrothermal waters in Yellowstone National Park, Wyoming, USA. II. Formation and decomposition of thiosulfate and polythionate in Cinder Pool. *J Volcanol Geoth Res* **97**: 407-423

Tables.

Table 3.1. Sample information

Sample	Geothermal Spring Group[a]	Temp °C	pH	Archaeal V3 reads	Bacterial V3 reads	Bacterial V1 Reads
APT001	Ampitheatre Springs	69.5	2.2	1553	977	7398
NMC002	Norris-Mammoth Corridor	42.3	2.14	4326	1902	
NOR003	One Hundred Spring Plain (NGB)	80.2	3.65	4283	2536	
NOR004	Back Basin (NGB)	88.4	5.76	5809	913	
NOR005	Back Basin (NGB)	70.1	8.15	1172	1557	
NOR006	One Hundred Spring Plain (NGB)	85.1	4.25	4856		
NOR007	One Hundred Spring Plain (NGB)	75.2	3.11	4352	3594	
NOR008	One Hundred Spring Plain (NGB)	63	3.01	4378	1511	
GIB009	Sylvan Springs (GGB)	68.3	2.39	5597	3021	
LCB011	Lower Culex Basin (LGB)	68.8	6.95	1813	2314	3814
LCB012	Lower Culex Basin (LGB)	59.9	3.99	6078	1812	
LCB013	Lower Culex Basin (LGB)	68.8	8.6	2644	2877	6087
MID014	Rabbit Creek (MGB)	72.5	6.01	673	2371	5480
MID016	Rabbit Creek (MGB)	41.5	3.06	2001	3437	6538
MID017	Rabbit Creek (MGB)	66.8	8.07	1070	3410	4169
MID018	Rabbit Creek	86	8.29	3843	3484	6047

	(MGB)					
LOW019	River (LGB)	53.7	8.83	658	2008	
LOW020	River (LGB)	52.8	5.32	2074	2705	5938
LOW021	River (LGB)	70	8.63	1178	1452	3594
LOW022	River (LGB)	81.8	8.12	6540	793	
LOW023	River (LGB)	73.5	5.93	5840	1371	8270
LOW035	Sentinel Meadows (LGB)	81.9	7.55	1402		5337
LOW037	Sentinel Meadows (LGB)	72	8.85	7546	2349	
SMH038	Seven Mile Hole	67	7.75	11728	4505	
SMH039	Seven Mile Hole	68.2	7.06	1876	3594	
LOW040	White Creek (LGB)	81.7	8.12	9940	750	
LOW041	White Creek (LGB)	85.7	8.26	6918	704	4839
LOW042	White Creek (LGB)	87.3	7.83	3216	1112	5800
LST043	Lone Star (LS)	42.4	6.01	2934	1201	7009
LST044	Lone Star (LS)	47.2	3.31	2616	3014	
MUD045	Mud Volcano	61.3	4.8	1204	907	
MID059	Rabbit Creek (MGB)	39	8.66	2532	1962	4771

[a]Geyser basins are given in parentheses are as follows: Norris (NGB), Gibbon (GGB), Lower (LGB), Midway (MGB), Lone Star (LS)

Table 3.2. OTU Diversity

Sample	Arc Chao1	Bac V3 Chao1	Bac V1 Chao1
APT001	138	18	1
NMC002	42	49	
NOR003	25	21	
NOR004	44	59	
NOR005	48	48	
NOR006	91		
NOR007	41	115	
NOR008	29	39	
GIB009	77	36	
LCB011	88	195	183
LCB012	24	82	
LCB013	175	113	49
MID014	100	107	110
MID016	38	74	64
MID017	47	115	138
MID018	51	34	59
LOW019	27	72	
LOW020	33	150	119
LOW021	223	103	114
LOW022	45	27	
LOW023	115	34	22
LOW035	20		34
LOW037	43	69	
SMH038	13	7	
SMH039	253	43	
LOW040	25	36	
LOW041	37	56	19
LOW042	32	34	33
LST043	282	201	111
LST044	35	53	
MUD045	56	38	
MID059	82	324	417

Table 3.3. Environmental correlates to community composition, pH and temperature

Correlate	Archaea[a]	Bacteria[a]	pH[b]	Temp[b]
pH	0.87***	0.70***	-	0.28
SO_4^{2-}	0.73***	0.49***	-0.82***	-0.30
Fe	0.67***	0.64***	-0.85***	-0.33
HCO_3^-	0.67***	0.43**	0.83***	0.27
Al	0.63***	0.64***	-0.82***	-0.15
Temperature	0.53***	0.48***	0.28	-
F^-	0.53***	0.47***	0.78***	0.16
Zn	0.41**	0.41***	-0.63***	0.06
Ba	0.38**	0.31**	-0.61***	-0.06
Na	0.35**	0.48***	0.65***	0.42*
Conductivity	0.27**	0.37**	0.00	0.36*
K	0.27*	0.29*	-0.09	0.39*
Si	0.26*	0.54***	0.21	0.36*
NO_3^-	0.26*	0.35**	-0.22	-0.15
B	n.s.	0.35**	0.38*	0.37*
As	n.s.	0.34**	0.50**	0.30
Br^-	n.s.	0.34**	0.30	0.42*
Cl^-	n.s.	0.30**	0.29	0.42*

[a]Correlation calculated by fitting variable to NMDS plot
[b]Correlation calculated using Pearson's product-moment
***Significant at the 0.001 alpha level
**Significant at the 0.01 alpha level
*Significant at the 0.05 alpha level

Table 3.4. OTU nodes with highest closeness centrality

OTU number	CC	Node Degree	Taxonomic Classification
11	0.64	67	Aigarchaeota; *Ca. Caldiarchaeum*
113	0.59	52	Chloroflexi; Unclassified
67	0.56	45	Thermi; *Thermus*
84	0.55	34	Aquificae; Aquificaceae
105	0.53	39	Armatimonadetes; OS-L
87	0.51	16	Proteobacteria; Thermodesulfobacteriales
100	0.50	32	Thermi; *Thermus*
121	0.50	37	Chloroflexi; *Chloroflexus*
88	0.49	18	Chlorobi; OPB56
8	0.49	18	Unassigned archaeon

Figures.

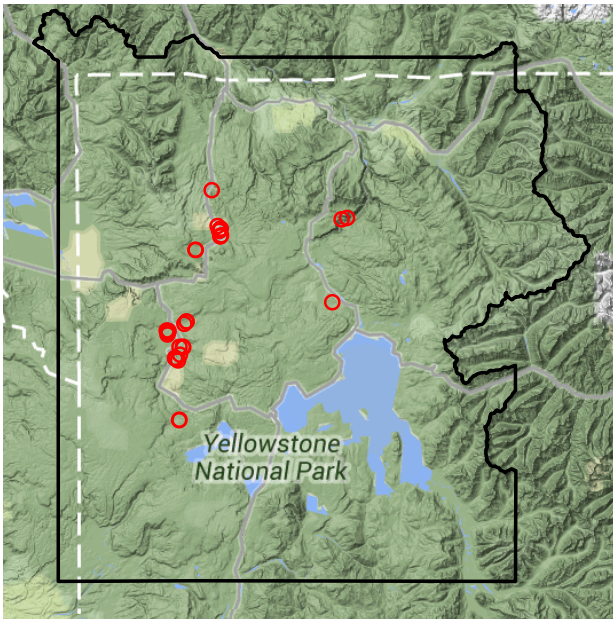


Figure 3.1. Sampling localities. Spring localities are overlaid in red circles on a topographic map of Yellowstone National Park. The border of the park is indicated by a solid black line.

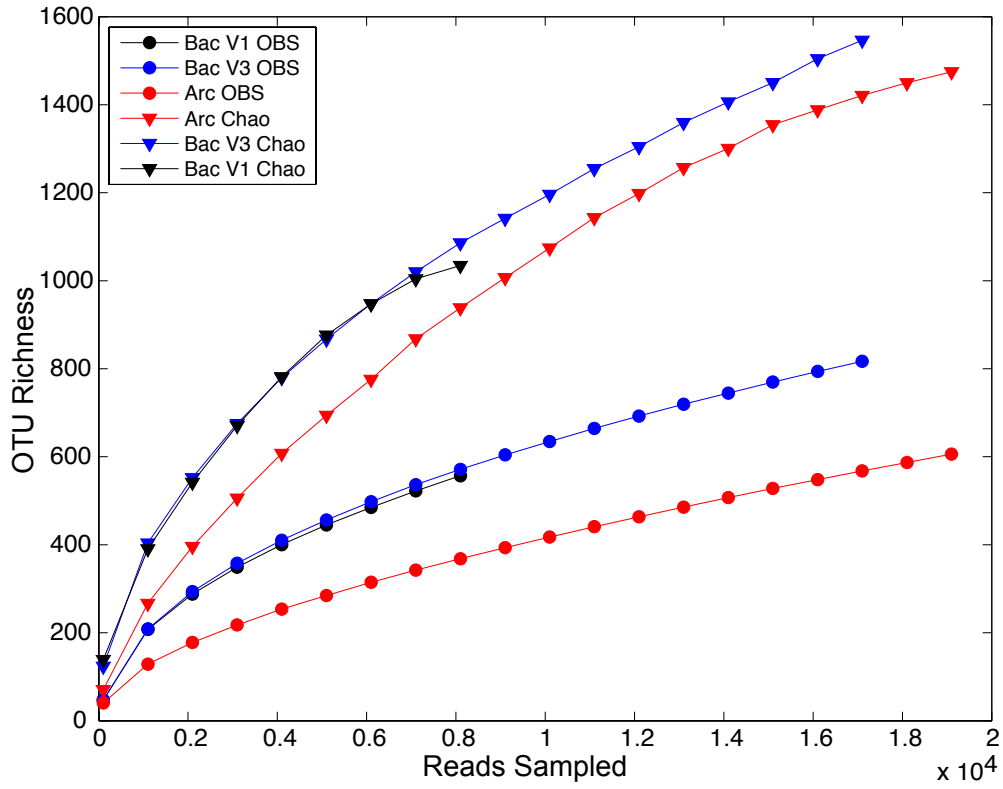


Figure 3.2. Rarefaction curves for archaeal and bacterial 16S rRNA gene diversity. Rarefaction curves for Chao1 estimated terminal richness and observed OTU richness (OBS) were calculated after subsampling each spring sample to 600 reads/sample. OTUs were defined at the 97% level similarity cutoff.

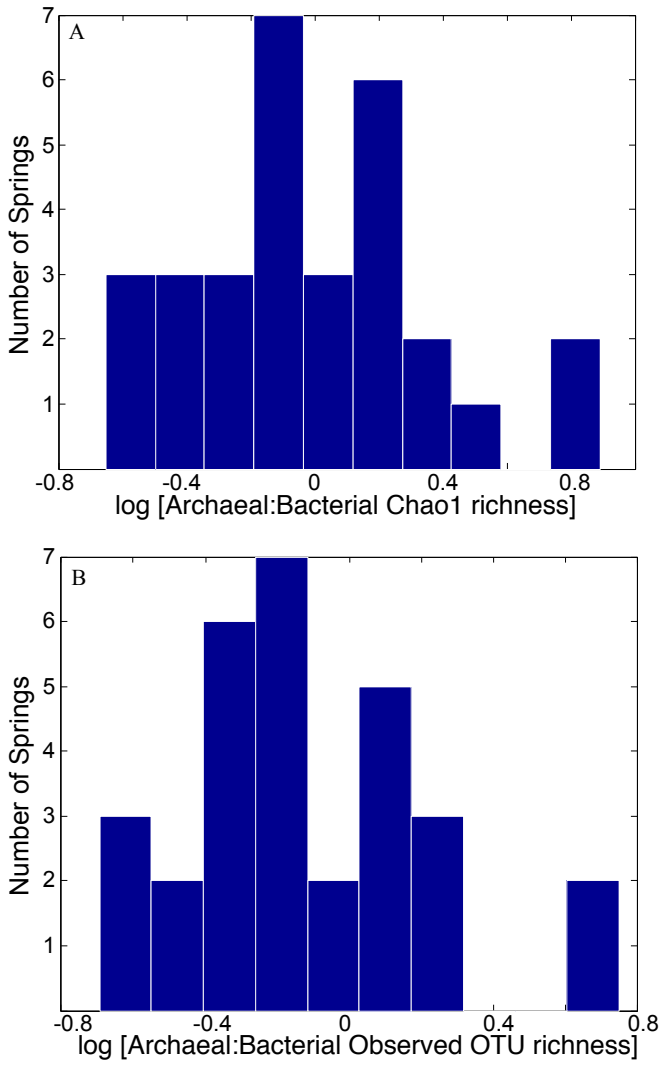


Figure 3.3. Histograms of the between-domain richness ratio based on the V3 region 16S rRNA gene richness estimates (A) The log transformed Chao1 richness estimate ratio for each spring. **(B)** The log transformed observed OTU richness ratio for each spring. Values for springs >0 indicate more archaeal-rich springs, while values <0 indicate more bacterial-rich springs. Both distributions were normally distributed (Shapiro-Wilk normality test $P > 0.05$ for both). The mean ratio for Chao1 estimate was not significantly different from 0 ($P > 0.05$, mean = -0.02, $t=-0.34$), while the observed

OTU ratio was skewed towards higher bacterial diversity ($P = 0.05$, mean = -0.13, t=-2.014).

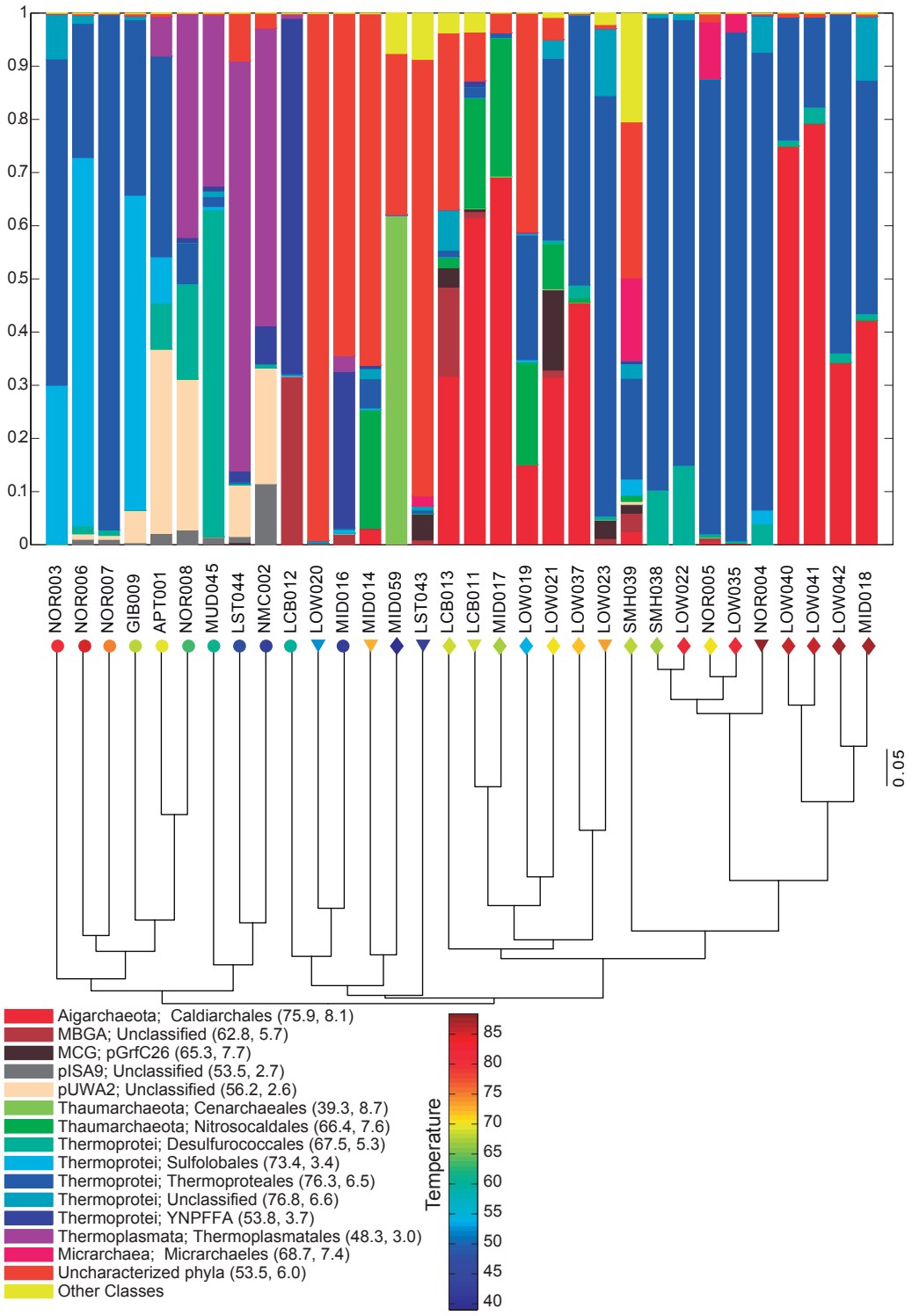


Figure 3.4. Archaeal community composition and dendrogram of community dissimilarity. Relative abundances are given for the 15 most overall abundant order-level

taxonomic groups. The abundance-weighted temperature and pH means for each taxonomic group is calculated as the weighted arithmetic mean for abundances (excluding sites where a lineage was not present) and given next to the name of each lineage. Symbols below each taxonomic distribution are colored by temperature according to the scale at the bottom right and are coded by low pH ($\leq$ 5; circles), slightly acidic (5-7; triangles) and circumneutral/alkaline ($\geq$7; diamonds). The dendogram was clustered using Bray-Curtis community distances.

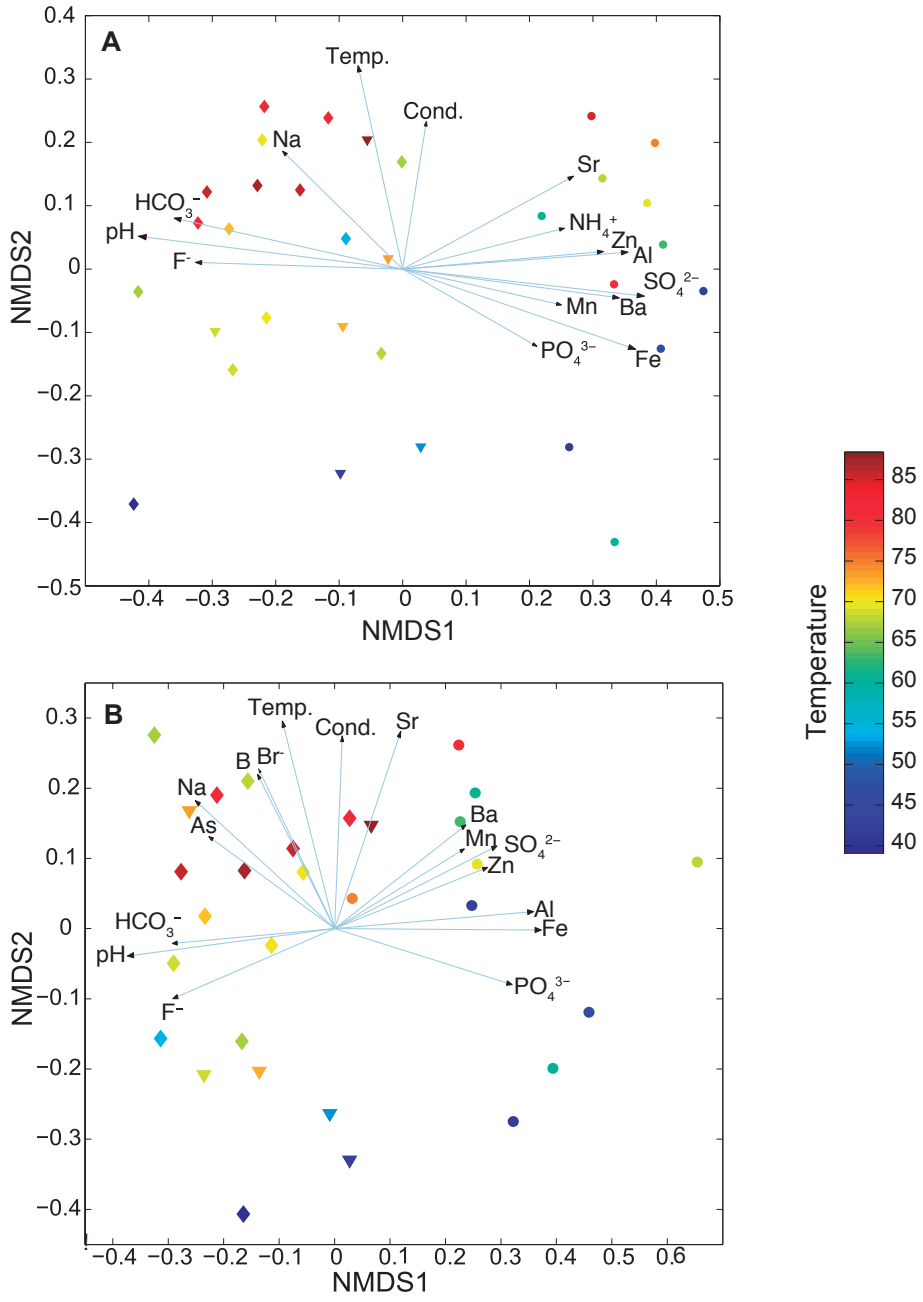
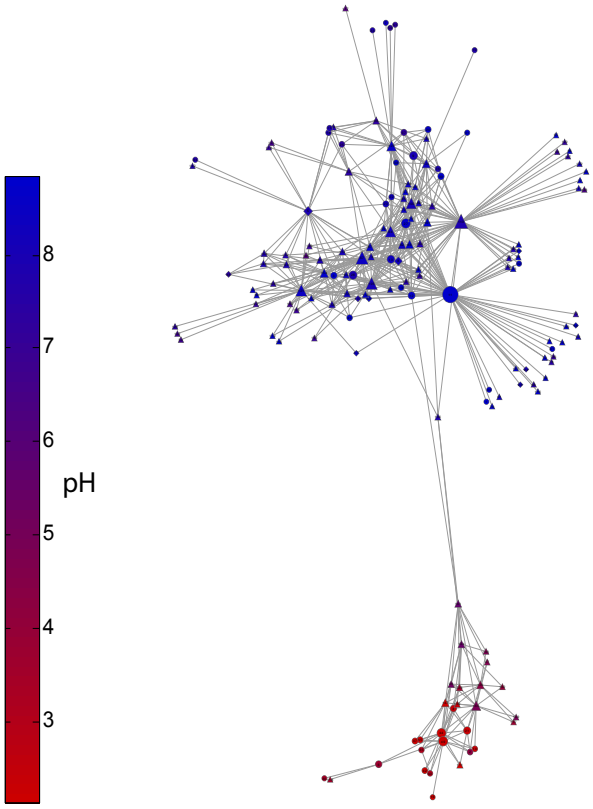


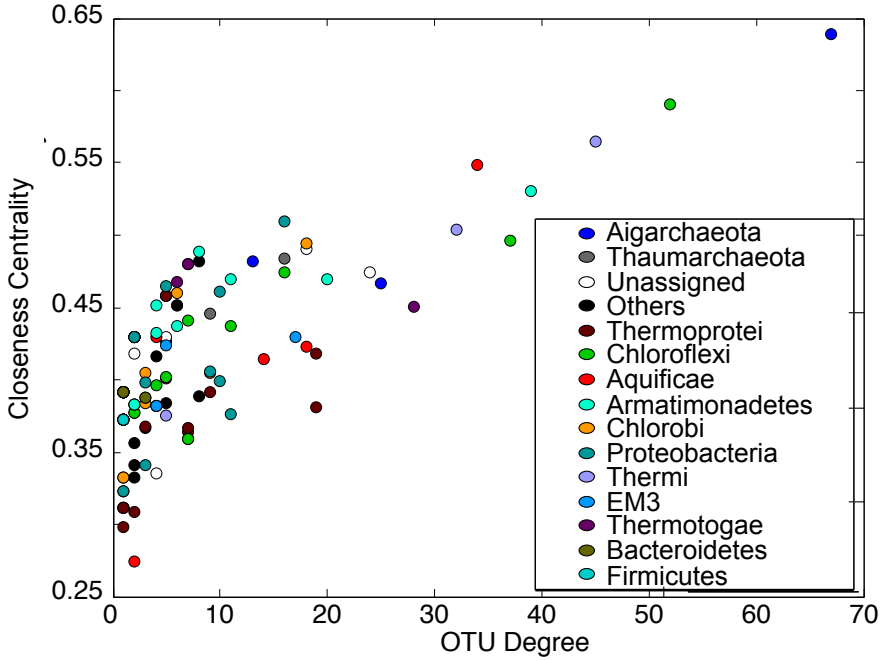
Figure 3.5. NMDS plots of archaeal and bacterial community composition and correlated environmental parameters. (A) Ordination of archaeal community composition and **(B)** bacterial community composition. Ordinations were performed using Bray-Curtis distances. Samples are colored according to spring sample temperature

according to the scale on the right. Sample symbols differ by pH of sites: low pH ($\leq$ 5 circles), slightly acidic (5-7; triangles), and circumneutral/alkaline ($\geq$ 7 diamonds). Significant ($P \leq 0.05$) correlations of environmental parameters to sample ordination are given by arrows from the origin, where the magnitude of the arrow is proportional to the correlation strength and the direction of the arrow is in the direction of highest correlation to the ordination.

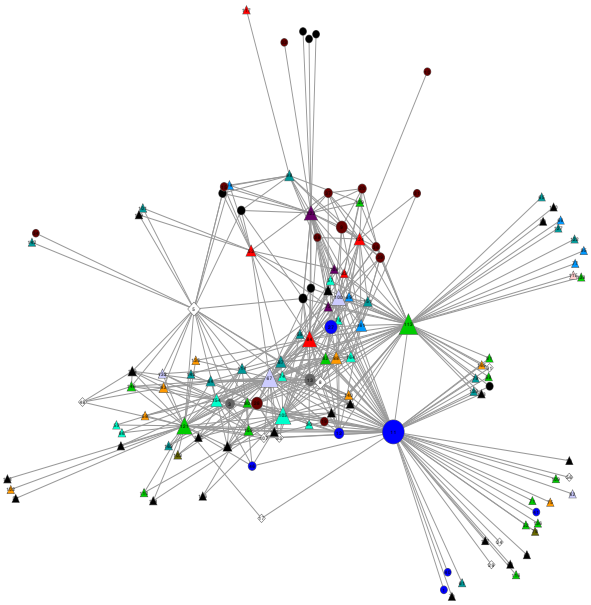
A.


8
7
6
pH
5
4
3

B.


0.65
0.55
0.45
0.35
0.25
Closeness Centrality
0
10
20
30
40
50
60
70
OTU Degree
Aigarchaeota
Thaumarchaeota
Unassigned
Others
Thermoprotei
Chloroflexi
Aquificae
Armatimonadetes
Chlorobi
Proteobacteria
Thermi
EM3
Thermotogae
Bacteroidetes
Firmicutes

C.



D.

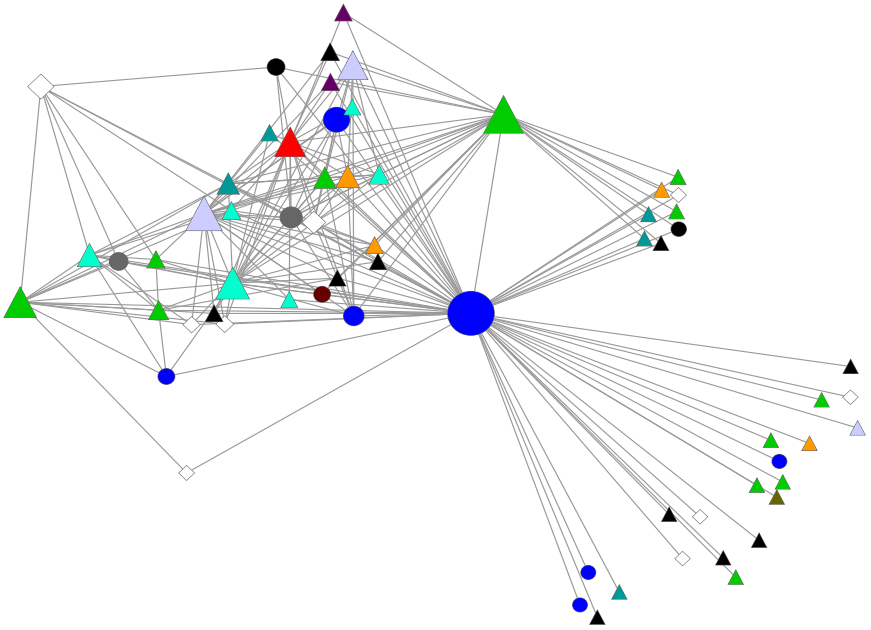
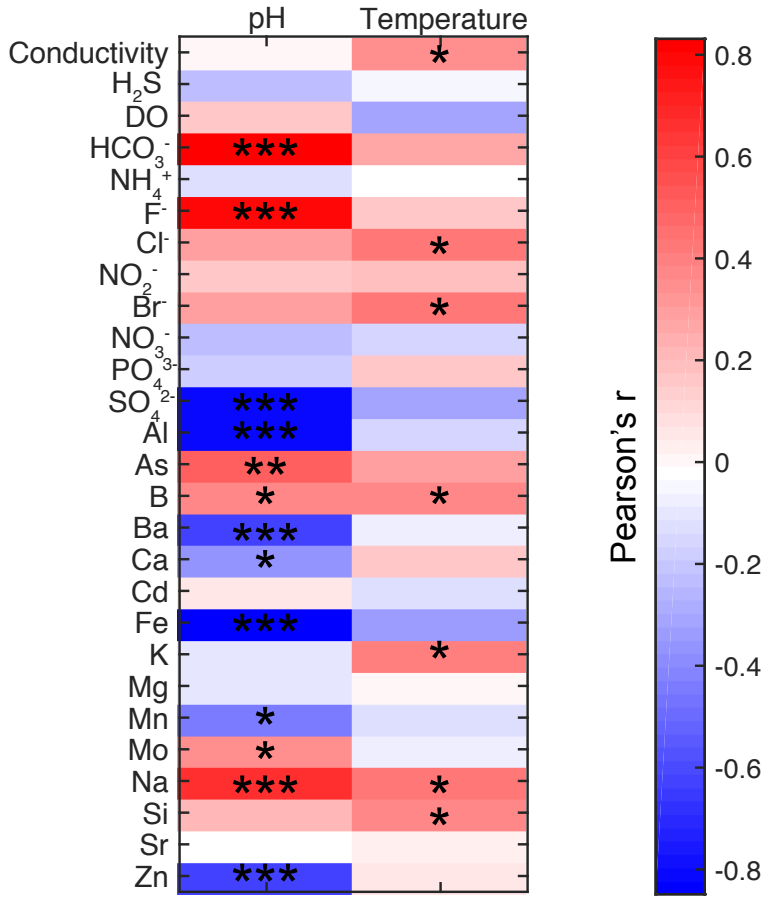


Figure 3.6. Co-occurrence network analysis. (A) Co-occurrence network diagram for all significant, positive OTU co-occurrences between domains. Each node represents an archaeal (circles) or bacterial (triangles) OTU and edges represent co-occurrences. Unclassified phyla are represented by diamonds. The nodes are arranged using the Edge-

Weighted Spring Embedded algorithm in Cytoscape. Nodes are colored according to the mean spring pH where they occur according to the scale on the left and node size is scaled to overall relative abundance. **(B)** Plot of closeness centrality (a quantitative measure of closeness to other nodes in the network) against the OTU degree for each node in the circum-neutral/alkaline subnetwork. Nodes are colored by their taxonomic classification. **(C)** Circumneutral/alkaline subnetwork of the total OTU network. Node symbols and colors are as in A & B. **(D)** Subnetwork consisting of the highest degree node (Aigarchaeota-affiliated) and other nodes directly connected to it. Node symbols and colors are as in A & B.

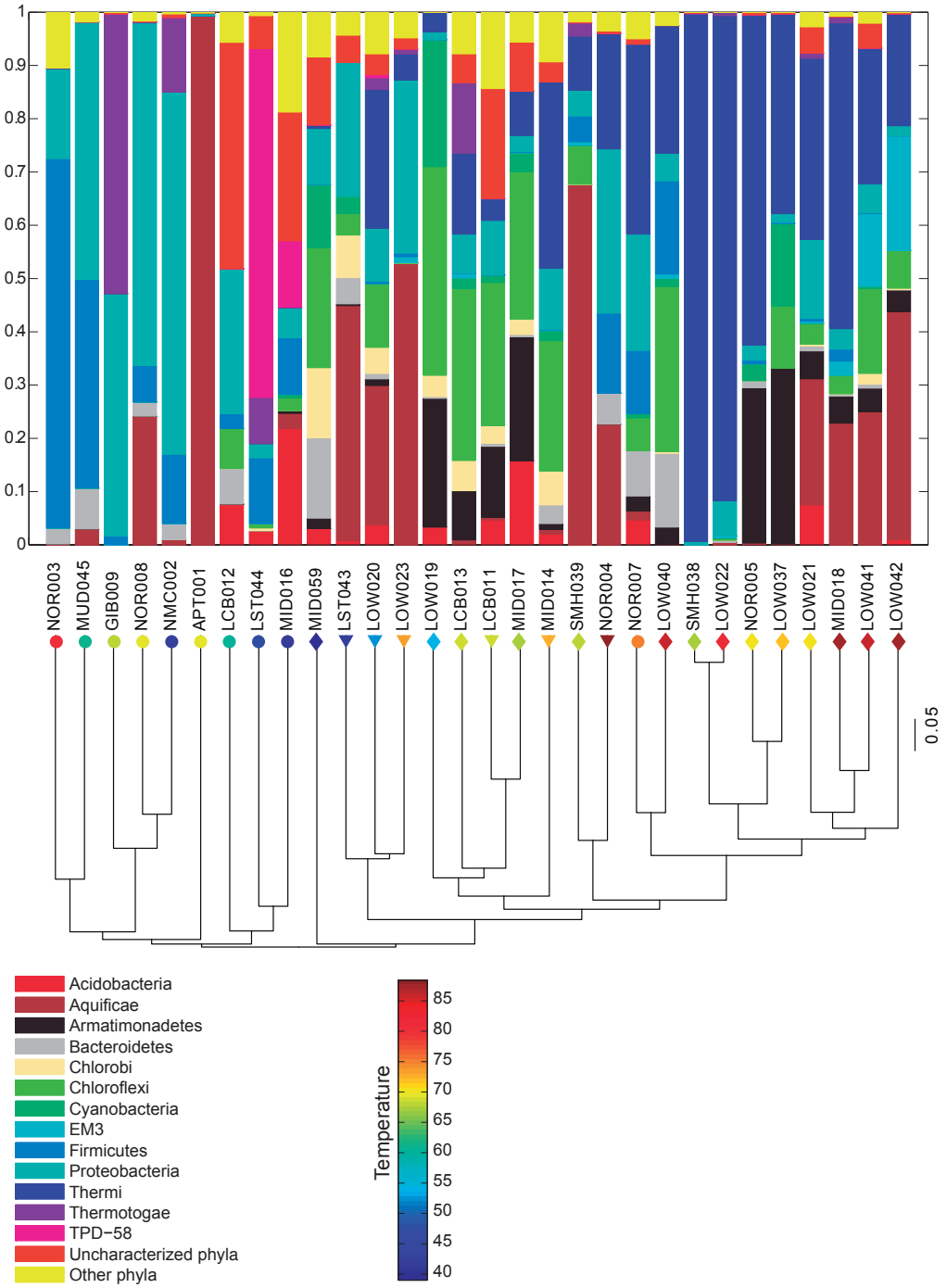
Supplementary Figures

Figure 3.S1. Heatmap of correlation between geochemical analytes against temperature and pH.



The strength of the Pearson's r correlation is given by the scale bar on the right. Statistically significant correlations are given by: * ($P \leq 0.05$), ** ($P \leq 0.01$) and *** ($P \leq 0.001$).

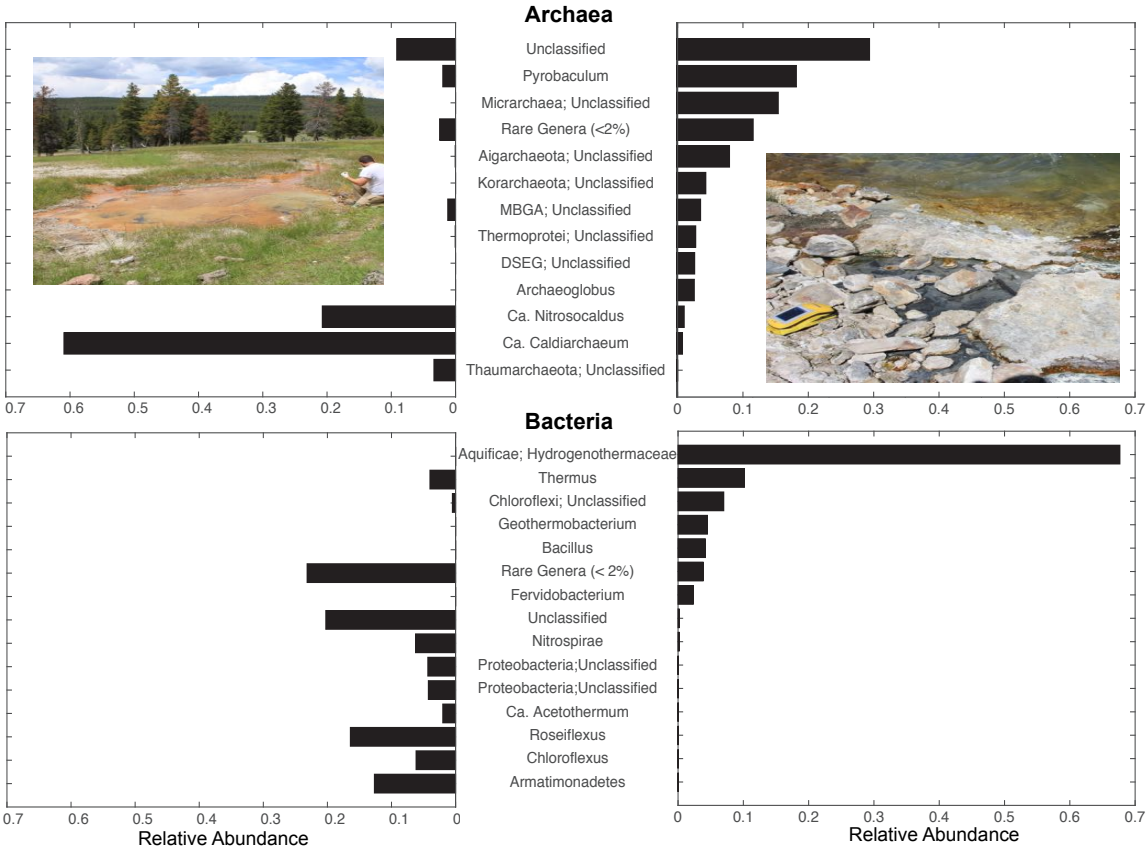
Figure 3.S2. Bacterial community composition and dendrogram of community dissimilarity.



Relative abundances are given for the 15 most abundant phylum-level taxonomic groups. Symbols below each taxonomic distribution are colored by temperature according to the

scale at the bottom right and are coded by pH: low pH (≤ 5; circles), slightly acidic (5-7; triangles) and circumneutral/alkaline (≥7; diamonds). The dendogram was clustered using Bray-Curtis community distances.

Figure 3.S3. Comparison of archaeal and bacterial community composition between springs LCB011 and SMH039.



Relative abundances of major genera in the springs is given for Archaea (top) and Bacteria (bottom) for two springs (LCB011 - left, SMH039 - right). Highest level of taxonomic classification is given for each. Inset into the archaeal relative abundance plots are images of each spring.

Chapter 4

Characterization of Novel, Deep-branching Heterotrophic Bacterial Populations Recovered from Thermal Spring Metagenomes

Daniel R. Colman[1], Zackary J. Jay[2], William P. Inskeep[2], Ryan deM. Jennings[2], Kendra R. Maas[3], Douglas B. Rusch[4] and Cristina D. Takacs-Vesbach[1]

[1]Department of Biology, University of New Mexico, Albuquerque, NM, USA; [2]Thermal Biology Institute and Department of Land Resources and Environmental Sciences, Montana State University, Bozeman MT; [3]Biotechnology-Bioservices Center, University of Connecticut, Storrs, CT; [4]Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana, USA

Submitted: *The International Society for Microbial Ecology Journal*

Abstract

Thermal spring ecosystems are a valuable resource for the discovery of novel hyperthermophilic *Bacteria* and *Archaea*, and harbor deeply-branching lineages that provide insight regarding the nature of early microbial life. We characterized bacterial populations in two circumneutral (pH ~ 8) Yellowstone National Park thermal (T ~ 80 ^{o}C) spring filamentous 'streamer' communities using metagenomic DNA sequence to investigate the metabolic potential of these novel populations. Four *de novo* assemblies representing three abundant, deeply-branching bacterial phylotypes were recovered. Analysis of conserved phylogenetic marker genes indicated that two of the phylotypes represent separate groups of an uncharacterized phylum (for which we propose the candidate phylum name 'Pyropristinus'). The third new phylotype falls within the proposed Calescamantes phylum. Metabolic reconstructions of the 'Pyropristinus' and

Calescamantes populations showed that these organisms appear to be chemoorganoheterotrophs, and have the genomic potential for aerobic respiration and oxidative phosphorylation. A survey of similar phylotypes (> 97% nt identity) within 16S rRNA gene datasets suggest that the newly described organisms are restricted to terrestrial thermal springs ranging from 70 - 90 ^{o}C and pH values of ~ 7 - 9. The characterization of these lineages is important for understanding the diversity of deeply-branching bacterial phyla, and their functional role in high-temperature circumneutral 'streamer' communities.

Introduction

The discovery and characterization of early-branching lineages of *Bacteria* and *Archaea* has been crucial for studying the origin and evolution of life on Earth. There is considerable evidence for the hypothesis that life originated in environments similar to modern hydrothermal settings, although other scenarios are also proposed (e.g. cold origins; Price 2009). Several findings support the hypothesis that life originated in thermal environments. Hyperthermophiles inhabit geothermal environments that are analogous to those of early Earth (Baross and Hoffman 1985), and are generally the deepest branching representatives of the tree of Life (Di Giulio 2003, Stetter 2006). The well-characterized and largely hyperthermophilic bacterial phyla *Aquificae* and *Thermotogae* have been considered the oldest bacterial lineages on the basis of phylogenetic evidence (Barion *et al.* 2007, Zhaxybayeva *et al.* 2009). More recently, an uncultured bacterium from subsurface thermal fluids, *Candidatus* 'Acetothermum autotrophicum', has also been posited as one of the earliest lineages in the *Bacteria* based on phylogenetic analysis of genome sequence (Takami *et al.* 2012). Consequently,

discovery and characterization of deep-branching lineages of thermophilic microorganisms is extremely useful toward the broader goal of understanding potential genomic and metabolic attributes of organisms most closely related to hyperthermophilic ancestors that were present during the earliest stages of bacterial evolution. Our understanding of the role of deep-branching hyperthermophiles in microbial evolution will benefit from a thorough description of modern day analogs to environments potentially similar to those important in the origin of life.

The characterization of uncultured populations of microorganisms from thermal environments has been integral for expanding the scope of known microbial diversity. Early phylogenetic surveys based on 16S rRNA gene populations revealed a significant diversity of uncultivated microorganisms in various hydrothermal settings, including numerous candidate phyla (Barns *et al.* 1994, Reysenbach *et al.* 1994, Hugenholtz *et al.* 1998, Takai and Horikoshi 1999). However, due to the difficulty of cultivating environmentally relevant microorganisms (particularly extremophiles), the physiologic diversity of many of these phyla has remained largely unknown since their discovery. Environmental genomics (e.g. metagenomics and single-cell genomics) has provided a valuable tool for assessing the metabolic capabilities and phylogenetic diversity of thermophiles and other extremophilic *Bacteria* and *Archaea* (Baker *et al.* 2010, Takami *et al.* 2012, Dodsworth *et al.* 2013, Kantor *et al.* 2013, Kozubal *et al.* 2013, Nunoura *et al.* 2011, Rinke *et al.* 2013, Hedlund *et al.* 2014). As a result of advances in environmental microbial genomics, major discoveries in microbial evolution and the ecology of uncultured populations have been documented (Baker *et al.* 2010, Nunoura *et al.* 2011, Inskeep *et al.* 2013, Rinke *et al.* 2013). However, there is still considerable microbial

diversity that remains uncharacterized and continued studies in high-temperature habitats will likely yield even greater resolution regarding microbial evolution, and the ecology of these environments.

Filamentous 'streamer' communities containing members of the *Aquificales* are common in geothermal outflow channels and hydrothermal vents in marine systems globally (Ferrera *et al.* 2007). We recently described and characterized metagenomes from six filamentous 'streamer' communities from geochemically distinct habitat types from Yellowstone National Park (YNP) (Inskeep *et al.* 2013, Takacs-Vesbach *et al.* 2013). The findings revealed that three primary genera of *Aquificales* dominate different streamer communities based on geochemical conditions (e.g., pH, sulfide), and that each habitat type supported different co-occurring community members, including several heterotrophic taxa. Two high-pH streamer communities (pH ~ 7.8; Octopus and Bechler Springs) contained abundant *Thermocrinis*-like populations (*Aquificales*) and several novel bacterial populations, which were not discussed in detail due to uncertainty in phylogenetic identification and lack of appropriate references available at the time of publication. A novel archaeal population from the Octopus Spring (OS) streamer communities was recently characterized (Beam *et al.,* submitted), and belongs to the candidate phylum Aigarchaeota; however, several novel and abundant bacteria in these communities have remained uncharacterized. Consequently, the objectives of this study were to 1) obtain *de novo* sequence assemblies corresponding to the uncharacterized three new bacterial phylotypes, 2) perform phylogenomic and functional analysis of the phylotype assemblies, and 3) determine the distribution of these populations in YNP and other thermal environments.

Here we describe nearly complete genomes from four *de novo* sequence assemblies (three phylotypes with one replicated in two springs) that were produced from random shotgun (Sanger) sequencing of two high-pH (~ 7.8) filamentous 'streamer' communities (temperature ~ 80 ^{o}C) from Octopus and Bechler Springs (Yellowstone National Park). These populations are representatives of two distinct deeply-branching candidate phylum-level lineages in the *Bacteria* ('Pyropristinus' is proposed here for one of the lineages containing two of the phylotypes, while the other novel phylotype belongs to the candidate Calescamantes phylum; Rinke *et al.* 2013). Metabolic reconstruction of these phylotypes indicated they are likely aerobic chemo-organoheterotrophs. Moreover, the distribution of 16S rRNA genes corresponding to these populations in YNP thermal spring databases suggested that all three are hyperthermophiles, and are found in terrestrial thermal springs, largely within circumneutral (pH ~ 7 - 8) filamentous 'streamer' communities. The discovery and characterization of these early-branching bacteria is critical for dissecting microbial community structure and function in modern-day high-temperature habitats, and provides significant opportunities for understanding the evolution of deeply-branching hyperthermal bacterial lineages.

Materials And Methods

Recovery of uncharacterized bacterial populations from hot-spring metagenomes

Site sampling, metagenome sequencing, assembly and annotation have all been described elsewhere (Inskeep *et al.* 2013, Takacs-Vesbach *et al.* 2013). Briefly, filamentous microbial community samples were collected near the top of the spring runoff channels from a spring in the Bechler Three Rivers Junction region of YNP (T = 80 - 82 ^{o}C, pH = 7.8) and Octopus Spring in the Lower Geyser Basin of YNP (T = 80 -

82 ^{o}C, pH = 7.9). A phenol/chloroform extraction method was used to extract community DNA (Inskeep *et al.* 2010), which was then used to construct a small-insert clone library. Sanger sequencing was used for random shotgun sequencing of the inserts (~40 Mb total DNA sequence for each site). Metagenomes were assembled using the Celera assembler and automated tools in the Integrated Microbial Genomes server (IMG; Markowitz *et al.* 2012) were used to predict and annotate genes. Nucleotide Word Frequency Principal Components Analysis (NWF-PCA) was used to identify predominant populations in the metagenomic contigs (> 2 kbp) as described previously (Takacs-Vesbach *et al.* 2013). The contigs were further analyzed and screened using G+C content (%) and phylogenetic analysis (most useful for phylotypes exhibiting closest neighbors above 80% nt ID) to obtain four *de novo* sequence assemblies corresponding to abundant and uncharacterized members of these communities (T1.1, T1.2, T2.1, T3.1; contig coverage > 1). Genome completeness was estimated using three metrics: tRNA synthetase complement was estimated by the presence of at minimum, one partial annotated gene for each of 21 prokaryotic genes coding for tRNA synthetases, the presence of 178 'conserved' bacterial housekeeping genes (Garcia Martin *et al.* 2006) and 40 'conserved' prokaryotic universal markers (Wu *et al.* 2013).

Phylogenetic relationships and placement in the bacterial phylogeny

Comparison by Amino Acid Identity

Amino Acid Identity (AAI) was calculated as the mean percentage of differing amino acid residues between homologous protein coding genes in pairwise comparisons of assemblies using blastp after filtering of low quality alignments (Konstantinidis and Tiedje 2005b). Protein coding gene alignments were used that 1) shared at least 30% a.a.

homology, 2) were alignable up to 70% of the length of the subject sequence, and 3) had an alignment length of at least 100 residues. The Calescamantes phylotype from OS (T3.1) was also compared to the recently characterized *Ca.* Calescibacterium nevadense (EM19-like) genome (IMG taxon ID: 2527291514). Average Nucleotide Identity (ANI) was calculated between the two closely related assemblies, T1.1 and T1.2 using default parameters with the online ANI calculator (http://enve-omics.ce.gatech.edu/ani/index; Goris *et al.* 2007).

Phylogenetic analysis

Phylogenetic analyses were conducted by surveying all three lineages (T1, T2 and T3) for homologous single-copy housekeeping genes (at least partial copies shared amongst all three lineages) that were previously identified as bacterial-specific and universal (archaeal/bacterial) marker genes (Wu *et al.* 2013). Genomic references were chosen based on blastp searches of ribosomal proteins against publically available genomes and curated such that every bacterial reference (Table 4.S1) contained $\geq$ 16 of the 18 total genes (5 archaeal/bacterial and 13 bacterial-specific marker genes; with the exception of *Ca.* C. nevadense; 3 genes missing). Each gene was aligned individually with Clustal Omega (Sievers *et al.* 2011), and alignment positions were confidence weighted to reduce the influence of ambiguously aligned positions using Zorro (Wu *et al.* 2012). An evolutionary model was chosen for each gene alignment using ProtTest v. 3.4 (Darriba *et al.* 2011). The concatenated gene alignment (8928 informative amino acid positions) was used in a maximum likelihood (ML) analysis in RAxML v. 7.3 (Stamatakis 2006) using alignment weights, and partitioning the concatenation so that each gene was modeled separately by the appropriate substitution model (primarily the

LG substitution model). Phylogenies were bootstrapped with 100 ML replicates. To assess the influence of alignment algorithm choice, the dataset was realigned using Muscle (Edgar 2004), analyzed as above, and the topology was compared to the Clustal aligned phylogeny.

Conserved Signature Indel (CSI) analyses were also used to assess if the lineages belonged to the phylogenetically closest related phyla, *Thermotogae* and *Aquificae*. CSIs specific to the *Thermotogae* (18 total; Gupta and Bhandari 2011) or *Aquificae* (4 total; Gupta and Lali 2013), relative to the rest of *Bacteria*, were used by referencing the assemblies against available *Thermotogae* and *Aquificae* genomes available in IMG. A total of 22 genes (encompassing 22 CSIs) were downloaded from IMG, aligned with Clustal, as described in the original publishing report, and inspected for the characteristic CSIs.

16S rRNA gene phylogenetic analysis was conducted using near full-length 16S rRNA genes (>1300 bp). The T3.1 and T1.1 assemblies were omitted from this analysis because they did not contain full-length 16S rRNA genes. Genes were aligned using PyNAST (Caporaso *et al.* 2010) with the Greengenes reference dataset (DeSantis *et al.* 2006). The DNA substitution model for the alignment was selected using Modeltest v. 3.7 (Posada and Crandall 1998) and the Akaike Information Criterion (AIC) model metric. ML analysis was conducted in MEGA v.6 (Tamura *et al.* 2013) using the General Time Reversible model with a proportion of invariant sites and Γ distribution of rates.

Metabolic Reconstruction

Annotated genes were used to assess the presence of metabolic pathways in all three lineages. The conspecific-level relatedness between T1.1 and T1.2 assemblies

allowed the use of the less complete T1.1 assembly to augment the genes not found in T1.2. Where possible, genome sequence of *Ca.* Calescibacterium nevadense (Rinke *et al.* 2013) was used to confirm the absence of pathways in T3, which was related to *Ca.* C. nevadense. Genomic data for the four assemblies produced here is available under the NCBI Bioproject ID PRJNA280379.

Ecological Distribution

16S rRNA genes of the three lineages were used in blastn searches against available datasets to determine the habitat distribution of these newly described populations. Because a full-length 16S rRNA gene was not present in the T3.1 assembly, a representative clone from the 16S rRNA gene library of the same Octopus Spring metagenome sample was used (Takacs-Vesbach *et al.* 2013). This clone group (EM19) was also described in pink-streamer communities of the same spring (Reysenbach *et al.* 1994; Blank *et al*., 2002). *Ca.* C. nevadense is also closely related to the EM19 clone from Octopus Spring (Rinke *et al.* 2013), and was the closest genome sequence data available for comparison to the Calescamantes-like assembly from Octopus Spring (Calescamantes-OS; T3.1 used here). Searches were conducted against 16S rRNA gene datasets including Genbank, IMG metagenomes, Greengenes (DeSantis *et al.* 2006), the Ribosomal Database Project (Cole *et al.* 2014), as well as YNP-specific surveys (including 454 pyrosequencing datasets) of 49 YNP springs spanning a wide range of temperature and pH values (Takacs-Vesbach *et al.* unpublished data) and clone-libraries of 82 YNP springs (Mitchell 2009). 16S rRNA gene matches with > 97% identity to each of the three lineages were considered a positive occurrence. Metadata for each reference sample (temperature, pH, and geographic location) were collected from the publishing

reports, where available, and augmented with data from the YNP Research Coordination Network database (http://www.rcn.montana.edu; Appendix File 4.A2). Mean values for sample temperature are used where ranges were reported. Statistical differences of temperature and pH distributions among groups were tested using a Kruskal-Wallis rank sum analysis of variance test in R (R Core Team 2014).

Results and Discussion

Recovery of uncharacterized bacterial populations from hot-spring metagenomes

Assembled metagenome sequence from Octopus and Bechler Springs was analyzed using nucleotide word frequency-principal components analysis (NWF-PCA) to obtain contigs and scaffolds (> 2kbp length only) sharing similar sequence character (Figure 4.1). These scaffolds and contigs were screened using G+C content (%) and phylogenetic analysis to obtain *de novo* sequence assemblies corresponding to each of the predominant phylotypes in these communities. Contig G+C content was highly uniform within assemblies after curation (Figure 4.S1). The Octopus Spring community contained at least eight predominant phylotypes (Desulfurococcales not shown), while Bechler Spring contained only three abundant phylotypes (Figure 4.1a). A detailed analysis of each of the predominant phylotypes showed that the two streamer communities contained highly-related populations of *Thermocrinis* spp. (order *Aquificales*), *Pyrobaculum* spp. (order *Thermoproteales*), and a novel Type 1 (T1) candidate phylum 'Pyropristinus' population. The streamer community from Octopus Spring contained abundant populations of an additional Type 2 (T2) 'Pyropristinus' population, and a close relative of the newly proposed bacterial phylum Calescamantes (formerly referred to as the EM19

candidate division; Rinke *et al.* 2013), an uncharacterized member of the Firmicutes, and a member of the candidate archaeal phylum Aigarchaeota (Beam *et al.,* submitted).

Random metagenome sequence reads (average read length = 820 bp) were reanalyzed based on nucleotide identity to the *de novo* assemblies built from these data sets (Figure 4.2). Consequently, phylogenetic deconvolution of the G+C (%) frequency plot shows the actual organisms represented by the assembled and characterized consensus sequence (shown at 90% nucleotide identity; Figure 4.2). The analysis of reads that share identity (> 90%) to individual *de novo* sequence assemblies provides a strong visual assessment of the community composition (Figure 4.2), and an excellent estimate of the abundance of these populations *in situ* (Table 4.S2). All three types ('Pyropristinus' Types 1 and 2 and the Calescamantes-like phylotype) were present in both springs, but were in greater abundances in Octopus Spring. Average estimates of genome completeness obtained from the presence of housekeeping genes in the *de novo* sequence assemblies based on three separate lists of genes (Table 4.1) were 65, 72 and 63% for the Type 1 (T1.2; Bechler), T2 and the Calescamantes-like populations, respectively (47% for T1.1; Octopus). Estimates by tRNA synthetase complement were higher (86% for both T1.2 and T2) than those based on the presence of 'conserved' housekeeping genes involved in many cellular processes (50-60% and 59-73% for T1.2 and T2, respectively). The lower estimates given by the presence of housekeeping genes involved in a variety of cellular processes may be conservative due to the lack of appropriate references for identifying these genes in novel, deep-branching organisms (discussed further below). The cumulative sequence and contig coverage plots, coupled with genome coverages of ~ 2 - 3.5x indicates that these genomes were adequately sampled (Figure 4.S1). Further, the

relative abundance of functional (COG) categories of genes is consistent with other closely related phyla, suggesting adequately sampled genomes for all four assemblies (Figure 4.S2).

Phylogenetic Analysis

Phylogenetic comparisons using concatenations of five universal genes across the *Bacteria* and *Archaea*, and 13 bacterial-specific single-copy marker genes suggest that the 'Pyropristinus' T1 and T2 lineages belong to a deeply-branching lineage of *Bacteria*, which is distinct from all currently characterized bacterial phyla (Figure 4.3). Analysis of only the shared ribosomal proteins (n=5) confirmed that the 'Pyropristinus' lineage is a well-supported clade, and is a basal group compared to the rest of *Bacteria* (Figure 4.S3). The phylogenomic analysis also confirmed that the ‘streamer’ community from Octopus Spring contained a population of the recently proposed Calescamantes phylum (formerly EM19; Rinke *et al.* 2013, Hedlund *et al.* 2014,). The T1.1 and T1.2 assemblies (from Octopus and Bechler Springs, respectively) were highly-related to one another independent of the comparison method: they exhibited ANIs of 96% $\pm$ 1.3% (n=2410), AAIs of 94% $\pm$ 10.1%, (n=743), and 16S rRNA gene identities of 98% (blastn). The high nucleotide and amino acid (a.a.) identities of the two T1 assemblies show that these populations belong to the same genus and likely to the same species (Konstantinidis and Tiedje 2005a, Konstantinidis and Tiedje 2005b).

The 'Pyropristinus' T1 and T2 assemblies however were substantially different (46.6% $\pm$ 12.3% AAI, n=442; Figure 4.4a), and each was also considerably different than the Calescamantes population from Octopus Spring (average AAI ~ 42%). While taxonomic rank delineations using AAI do not follow discrete cutoffs, an AAI of only

47% between T1 and T2 is consistent with phylum- or class-level differentiation (Konstantinidis and Tiedje 2005b). Further, 16S rRNA genes from T1.2 and T2 only shared 84% homology, which is also consistent with existing phylum- or class-level delineations (Yarza *et al.* 2014). Genomic differences between either the T1 or T2 populations and the Calescamantes-OS population were also consistent with phylum level differentiation (Konstantinidis and Tiedje 2005b), and supported the phylogenetic divergence of these newly described organisms. The Calescamantes population from OS was more closely related to the recently described candidate species *Ca.* Calescibacterium nevadense (78.0% $\pm$ 18.1% mean AAI, n=1053; Figure 4.4b) obtained from single cell amplified genomes (SAGs) from Great Boiling Spring, NV, USA (Rinke *et al.* 2013), yet different enough to suggest that these populations are likely different genera or families within the proposed Calescamantes phylum (Rinke *et al.* 2013). Phylogenetic analysis without archaeal outgroups confirmed a highly supported 'Pyropristinus'/Calescamantes group separate from the Aquificae/Thermodesulfobacteria clade, and which together are distinct from all other bacterial phyla (Figure 4.S4). Alignment of the concatenated gene dataset with an alternative algorithm (Muscle) recovered the strongly supported basal branch of 'Pyropristinus' and the Calescamantes relative to other *Bacteria*, and a nearly identical topology to the Clustal aligned dataset (data not shown).

Phylogenetic analyses of these bacteria using long-fragment 16S rRNA gene sequences recovered from the *de novo* sequence assemblies and/or environmental clones (largely from YNP) also showed that these organisms form a deep-branching group inclusive of both 'Pyropristinus' T1 and T2 populations, near the *Thermotogae*, and other

uncultured *Bacteria* (Figure 4.5). The T1 population is closely-related (98% 16S rRNA gene identity) to the uncharacterized EM3 bacterium originally discovered in Octopus Spring (Reysenbach *et al.* 1994). No substantial EM3 genomic references were available for phylogenomic comparisons. Partial genome sequence for a member of the EM3 lineage was recovered from single cell genomes from Great Boiling Spring, NV (14% estimated completeness by tRNA synthetase complement, IMG taxon ID: 2264867090; Rinke *et al.* 2013), but was not sufficiently complete for phylogenomic comparisons (e.g. Figure 4.3). The T1 and T2 lineages belonged to separate 16S rRNA gene clades, which is consistent with results from AAI comparisons, and suggests that T1 and T2 belong to different groups within the candidate 'Pyropristinus' phylum. 'Pyropristinus' types T1 and T2 formed a cohesive group with other uncultured organisms primarily from hydrothermal systems (mean 16S rRNA gene distance within the group = 16%), and which excluded other uncultured organisms largely from engineered water treatment reactor communities ('Uncultured Thermotogae-Like Group').

The lack of Conserved Signature Indels (CSIs) typical of *Thermotogae* and *Aquificae* genomes further supported the separation of T1, T2 and the Calescamantes phyla as distinct phyla. Of the 18 genes containing previously published *Thermotogae*-specific CSIs (Gupta and Bhandari 2011), 12 of the genes were present in at least one of the four assemblies. Of those twelve genes, ten of the *Thermotogae*-specific CSIs were not present in the assemblies (Appendix File 4.A3). Most notable were a 10-15 a.a. insertion in the 50S ribosomal protein L4 (rpL4) of *Thermotogae* that was not present in the 'Pyropristinus' T2 or Calescamantes phylotypes and a three a.a. insertion in the 50S ribosomal protein L7/L12 of *Thermotogae* also not present in any of the four assemblies

(Figure 4.S5). Of the four genes with previously published *Aquificae*-specific CSIs (Gupta and Lali 2013), two were not in any of the four assemblies and the two remaining provided conflicting results amongst the lineages. The single a.a. deletion in the ribosomal small subunit methyltransferase H (RsmH) of all *Aquificae* was also present in both T1 assemblies, but not in the T2 assembly (Figure 4.5). In addition, the T1.2 assembly contained a two a.a. insertion in the 50S ribosomal protein L15, that was not present in the Calescamantes assembly (Appendix File 4.A3). The CSI results, coupled with the 16S rRNA and phylogenomic analysis supports the differentiation of the 'Pyropristinus' lineage and the Calescamantes phylum from members of the Thermotogae and/or the Aquificae, which are the closest characterized relatives of the reported assemblies. The relationship of the 'Pyropristinus' and Calescamantes lineages to the recently described, deep-branching bacterium *Ca.* Acetothermum autotrophicum (Takami *et al.* 2012) was also attempted. However, due to a lack of universal housekeeping marker genes in the available sequence for *Ca.* 'A. autotrophicum' (only three universal markers were shared among *Ca.* A. autotrophicum: IMG taxon ID: 2540341180, T1 and T2), consistent and well-supported placement of *Ca.* A. autotrophicum relative to the 'Pyropristinus', Calescamantes, *Thermotogae* and *Aquificae* lineages could not be adequately assessed. A more robust set of universal marker genes from additional genome references will be necessary to confidently confirm the phylogenetic placement of 'Acetothermia'-like populations.

Metabolic Reconstruction and Potential Community Interactions

Central Carbon Metabolism

Metabolic reconstruction of the 'Pyropristinus' T1 and T2 populations showed that these organisms shared nearly all major biochemical attributes, despite their phylogenetic dissimilarity. Statistical analysis of the COG distributions from the 'Pyropristinus' (T1 and T2) and Calescamantes populations with *Aquificae*, *Thermodesulfobacteria* and *Thermotogae* references indicated that the functional content of the T2 assembly was very similar to the two T1 assemblies (Figure 4.S6). No evidence was found for inorganic carbon fixation pathways (Fuchs 2011) in either the 'Pyropristinus' (T1 and T2) or Calescamantes-OS populations. The lack of inorganic carbon fixation pathways in the Calescamantes-OS is consistent with analysis of the related *Ca.* C. nevadense (Hedlund *et al.* 2014).

The metabolism of polysaccharides was indicated in the T1, T2 and Calescamantes lineages by the presence of β-glucosidases and α-amylases, as well as other important protein coding genes in starch degradation (cellulase in T1; α-glucosidase and starch synthase in Calescamantes). An oligosaccharide transporter present in T1 also suggests that they may be utilizing exogenous saccharides produced by other autotrophic streamer community members, such as *Thermocrinis* spp. (*Aquificales*), or Aigarchaeota that are also present in these communities (Takacs-Vesbach *et al.* 2013; Beam *et al.* submitted). All genes necessary for the Embden-Meyerhoff glycolysis pathway were present in 'Pyropristinus' T1, and most were also present in the Calescamantes population (and *Ca.* C. nevadense) indicating the potential to oxidize glucose. The presence of an archaeal-like fructose1,6-bisphosphatase (*fbp*) also indicated that gluconeogenesis may occur via a bacterial variant of the bifunctional enzyme that is conserved in *Archaea* and early-branching bacterial lineages (Say and Fuchs 2010). A nearly complete TCA cycle

was also present in T1 (exclusive of *idh*) and both Calescamantes populations (Calescamantes-OS and *Ca.* C. nevadense). No evidence of anaerobic fermentation was found in the 'Pyropristinus' T1 populations; genes coding for proteins involved in acid or alcohol fermentation pathways including alcohol dehydrogenase, acetate kinase, formate dehydrogenases and associated hydrogenases were absent from the Type 1 and Type 2 assemblies. Alcohol dehydrogenases were present in the Calescamantes group (both the OS and *Ca.* C. nevadense assemblies), which suggests possible fermentation in those phylotypes. Both the T1 and Calescamantes groups also contained protein-coding genes involved in the oxidation of fatty acids to acetyl-CoA (β-oxidation pathway). Moreover, long-chain fatty acid transporters present in Type 1 populations may indicate dependence on fatty acids from other streamer community members for heterotrophic metabolism. Most amino acid synthesis pathways were observed in 'Pyropristinus' T1, with the exception of tryptophan, methionine, serine and threonine. The Calescamantes (OS and *Ca.* C. nevadense) genomes only lacked evidence for a glutamate synthesis pathway. Both groups contained several amino acid/peptide transporters, peptidases and proteases suggesting the ability to import oligopeptides, amino acids, and utilize peptides that may be present in the streamer microenvironment. 'Pyropristinus' and Calescamantes populations contained the necessary enzymatic pathways for the non-oxidative pentose phosphate pathways used in the generation of five carbon sugars for nucleotide synthesis, in addition to most of the genes necessary for purine and pyrimidine biosynthesis. Both lineages also contained evidence for NAD/P, riboflavin, pantothenate/Coenzyme A, folate, and pyridoxine biosynthesis pathways.

Energy Conservation

Nearly complete respiratory complexes including subunit I heme Cu oxidases were recovered in the T1 and T2 populations as well as the Calescamantes representatives, which strongly suggests that these organisms utilize oxygen for respiration and conduct oxidative phosphorylation. Key genes involved in NH_4^+ oxidation (*amo*), sulfur oxidation (*sqr*, *hdr*, *tqo*), sulfur/sulfate reduction (*psr*, *dsr*), H_2 oxidation (*hyn*), methanotrophy (*pmo*), arsenate/arsenite metabolism (*arr, aox*), and NO_3^- reduction (*nar, nap*) were not present in the 'Pyropristinus' populations or the Calescamantes-OS. An *sqr*-like gene present in the *Ca.* C. nevadense assembly suggests that HS^- may serve as an electron donor in that phylotype. A nitrite reductase (*nirS*) present in *Ca.* C. nevadense with high homology (70%) to the cytochrome cd_1 nitrite reductase from *Hydrogenobacter thermophilus* TK-6 (*Aquificales*) (Suzuki *et al.* 2006) along with *nosZ* nitrous oxide reductase genes in the OS/*Ca.* C. nevadense assemblies suggests the potential for dissimilatory nitrite reduction. Intriguingly, both *Ca.* C. nevadense and the Calescamantes-OS assembly contain *nosZ* genes with higher homology to Chloroflexi/Bacteroidetes *nosZ* (~63 and 64% *Caldilinea aerophilus* STL-6-01/ *Rhodothermus marinus* R-10; IMG gene IDs: 2540573393, 646411298, respectively) than to the nosZ of *H. thermophilus* (39-43%; IMG gene ID: 646540242). The inconsistent phylogenetic affiliation of denitrification subunits suggests divergent evolutionary histories for the genes involved in this pathway in the Calescamantes and *Aquificae*.

The 'Pyropristinus' T1 lineage and Calescamantes (OS/ *Ca.* C. nevadense) assemblies contained nearly complete NADH:quinone oxidoreductase (*nuo*) complexes necessary for NADH-mediated oxidative phosphorylation (Figure 4.6), but differed

significantly in key energy conservation mechanisms. The T1/T2 assemblies contained an archaeal/V (vacuolar)-type ATPase, while the Calescamantes population had a largely-complete F_0F_1 F-type ATPase complex. Only a small number of *Bacteria* contain archaeal V-type ATPases, whereas the F-type ATPase is ubiquitous and phylogenetically conserved among *Bacteria*, and is thought to be the ancestral bacterial ATPase (Mulkidjanian *et al.* 2007). The *Thermotogae* variously contain V-type and/or F-type ATPases (Nelson *et al.* 1999, Iida *et al.* 2002, Nesbo *et al.* 2002), whereas the *Aquificae* contain F-type ATPases (Koumandou and Kossida 2014). The recently described deep-branching bacterium, *Ca.* Acetothermum autotrophicum, also contains a V-type ATPase (Takami *et al.* 2012). The disparity in ATPase complexes between *Ca.* A. autotrophicum/*Thermotogae*/'Pyropristinus' lineages relative to the Calescamantes and *Aquificae* suggests a major divergence in energy conserving mechanisms among these phyla, and warrants further study of their evolutionary history.

Differences in Secondary Metabolites

Several differences among the 'Pyropristinus' and Calescamantes populations were also noted in secondary metabolite synthesis pathways. For example, T1 lacked all genes necessary for the synthesis of cobalamin (Vitamin B_{12}), whereas the Calescamantes-OS populations contained only 33% of the ~30 genes necessary for *de novo* synthesis (similar to *Ca.* C. nevadense, which contained ~53% of these genes). 'Pyropristinus' T1 and T2 populations contained genes coding for outer membrane cobalamin receptor proteins and a permease involved in cobalamin transport, which suggests that they likely import this cofactor from the environment. Vitamin B_{12} is a necessary cofactor for methylmalonyl-CoA mutase (*mcm*; present in T1 and in

Calescamantes), which is involved in the degradation of amino acids and fatty acids into succinyl-CoA (Martens *et al.* 2002). Consequently, the apparent reliance on vitamin B_{12} through different acquisition strategies may indicate a reliance on amino acids and fatty acids for energy conservation in both the 'Pyropristinus' and the Calescamantes populations. The presence of biotin synthase (*bioB*) as well as *bioADF* suggest that the Calescamantes populations are capable of synthesizing biotin; conversely, the presence of only one biotin synthesis gene (*bioH*) from the 'Pyropristinus' T1 lineages suggests potential biotin auxotrophy. All of the assemblies contain acetyl-CoA carboxylases, which require biotin for the synthesis of malonyl-CoA from acetyl-CoA in fatty acid synthesis (Streit and Entcheva 2003), and further supports a divergence in secondary metabolite acquisition between 'Pyropristinus' and the Calescamantes.

The 'Pyropristinus' assemblies lacked all genes necessary for flagellar synthesis, whereas the Calescamantes populations contained numerous flagellar biosynthesis genes including *flhA*, *fliM*, *fliN*, *fliE* and *flgC*. The *Ca.* C. nevadense genome contained many of the missing flagellar biosynthesis genes not observed in the Calescamantes population from OS, and suggests that they are both capable of flagellar-mediated motility. Chemotaxis genes *cheY* and *cheD* were present in the Calescamantes-OS population, whereas T1 contained *cheB*, *cheY* and *cheC*. Both the Calescamantes and 'Pyropristinus' populations appear to be gram negative based on the presence of the essential outer membrane protein assembly gene *yfiO* in Calescamantes, the *yaeT* outer membrane assembly gene in T1, and several other outer membrane associated proteins in both lineages (Bos *et al.* 2007, Sutcliffe 2011).

Ecological Distribution

Prior datasets of 16S rRNA gene diversity in YNP were queried for the presence of 'Pyropristinus' Types 1 and 2, and Calescamantes-OS populations. The presence of similar populations (> 97% 16S rRNA gene identity) is currently restricted to terrestrial thermal springs, largely in affiliation with *Aquificales* 'streamer' communities (Appendix File 4.A2). No representatives were found in marine hydrothermal settings. Moreover, these populations were only detected in high-temperature and circumneutral (pH ~ 6 - 9) geothermal springs, and only one Type 2-like phylotype has been observed outside of YNP (Figure 4.7; Appendix File 4.A2). The temperature and pH range of sites used to infer phylotype distribution (Takacs-Vesbach, unpl.) was statistically highly similar to the range observed for thermal springs within the entire YNP ecosystem (Pearson's $r = 0.66$, $P < 0.05$; Figure 4.S7), which suggests that this dataset was appropriate for inferring the presence or absence of these three populations with respect to temperature and pH within YNP. The observed temperature and pH ranges for 'Pyropristinus' T1, T2 and the Calescamantes-OS phylotypes were not significantly different from one another ($P > 0.05$), which suggests that they all occupy similar physicochemical niches. These results are also consistent with earlier observations of closely related EM3 and EM19-like populations in high-temperature circumneutral *Aquificae* dominated communities (Reysenbach *et al*. 1994, Blank *et al.* 2002, Meyer-Dombard *et al.* 2011).

Results from phylogenetic analyses show that two new phylotypes (Type 1 and Type 2) represent different groups of a phylum-level lineage distinct from other characterized bacterial phyla, and for which we propose the candidate genera epithets *'Candidatus* Caldipriscus sp. T1' (Cal'di.pris.cus. L. masc. adj. *caldus,* hot; L. masc. n. *priscus*, ancient or primitive; ancient thermophile) and '*Candidatus* Thermoproauctor sp.

T2' (Ther.mo.pro.auc'tor. Gr. fem. n. *therme*, heat; L. masc. n. *proauctor*, ancestor/founder; thermophilic ancestor), respectively. Further, on the basis of phylogenetic evidence, we propose the candidate phylum-level name 'Pyropristinus' (Pyr.o'pris.tin.us. Gr. neutr. n. *pyr*, fire; L. masc. adj. *pristinus*, former/early; early thermophiles) to include the *Ca*. Caldipriscus, *Ca.* Thermoproauctor and other closely related uncultured organisms, inclusive of the formerly identified EM3. The discovery and characterization of these new phylotypes will contribute to understanding the evolutionary relationships of deeply diverging *Bacteria*. The 'Pyropristinus' and Calescamantes populations described here appear to be reliant on carbon sources from other autotrophic members of the ‘streamer’ communities (and/or DOC present in spring water), and use reduced sources of organic carbon to respire aerobically. The consistency with which the 'Pyropristinus' and Calescamantes lineages co-occur with *Aquificae* in streamer environments suggests that these early branching bacteria may have co-evolved in circumneutral high-temperature environments. Differences in energy conservation mechanisms between the 'Pyropristinus' and Calescamantes lineages (e.g. ATPases, potential to respire anaerobically) suggests that they likely occupy different microenvironments across an oxygen gradient. Importantly, these newly-described phylotypes provide increased resolution of the metabolic attributes associated with deep-branching thermophilic bacterial lineages.

Acknowledgements

We thank Christie Hendrix and Stacey Gunther (Center for Resources) at Yellowstone National Park for research permitting and field assistance. Authors from MSU appreciate support from the Department of Energy (DOE)-Joint Genome Institute

Community Sequencing Program (CSP 787081), the DOE-Pacific Northwest National Laboratory (Foundational Science Focus Area) (MSU subcontract 112443) and the NSF IGERT Program (DGE 0654336). The work conducted by the Joint Genome Institute (DE-AC02-05CH11231) and the Pacific Northwest National Laboratory (Foundational Scientific Focus Area) is supported by the Genomic Science Program, Office of Biological and Environmental Research, US DOE. The UNM authors appreciate support from an NSF Biotic Surveys and Inventories grant (02-06773), the New Mexico Space Grant Consortium, the Louis Stokes Alliance for Minority Participation grant (NSF HRD 0832947), a UNM Office of Graduate Studies Graduate Student Success Scholarship, a UNM Office of Graduate Studies Research Project and Travel Grant and the UNM Research Allocation Committee.

References

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al.* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci U S A* **107:** 8806-8811.

Barion S, Franchi M, Gallori E, Di Giulio M (2007). The first lines of divergence in the Bacteria domain were the hyperthermophilic organisms, the Thermotogales and the Aquificales, and not the mesophilic Planctomycetales. *Bio Systems* **87:** 13-19.

Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci U S A* **91:** 1609-1613.

Baross JA, Hoffman SE (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins Life Evol B* **15:** 327-345.

Blank CE, Cady SL, Pace NR (2002). Microbial Composition of Near-Boiling Silica-Depositing Thermal Springs Throughout Yellowstone National Park. *Appl Environ Microbiol* **68:** 5123-5135.

Bos MP, Robert V, Tommassen J (2007). Biogenesis of the Gram-Negative Bacterial Outer Membrane. *Annu Rev Microbiol* **61:** 191-214.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26:** 266-267.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y *et al.*(2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42:** D633-642.

Darriba D, Taboada GL, Doallo R, Posada D (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27:** 1164-1165.

DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM *et al.*(2006). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34:** W394-W399.

Di Giulio M (2003). The Universal Ancestor and the Ancestor of Bacteria Were Hyperthermophiles. *J Mol Evol* **57:** 721-730.

Dodsworth JA, Blainey PC, Murugapiran SK, Swingley WD, Ross CA, Tringe SG *et al.*(2013). Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat Commun* **4:** 1854.

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

Ferrera I, Longhorn S, Banta AB, Liu Y, Preston D, Reysenbach AL (2007). Diversity of 16S rRNA gene, ITS region and aclB gene of the Aquificales. *Extremophiles* **11:** 57-64.

Fuchs G (2011). Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life? *Annu Rev Microbiol* **65:** 631-658.

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC *et al.*(2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24:** 1263-1269.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57:** 81-91.

Gupta RS, Bhandari V (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. *Antonie van Leeuwenhoek* **100:** 1-34.

Gupta RS, Lali R (2013). Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order Aquificales, containing the families Aquificaceae and Hydrogenothermaceae, and a new order

Desulfurobacteriales ord. nov., containing the family Desulfurobacteriaceae. *Antonie van Leeuwenhoek* **104:** 349-368.

Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles* **18:** 865-875.

Huber H, Prangishvaili D (2006). Sulfolobales. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

Huber R, Eder W (2006). Aquificales. In: Dworkin M, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York.

Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998). Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* **180:** 366-376.

Iida T, Inatomi K, Kamagata Y, Maruyama T (2002). F- and V-type ATPases in the hyperthermophilic bacterium *Thermotoga neapolitana*. *Extremophiles* **6:** 369-375.

Inskeep WP, Rusch DB, Jay ZJ, Herrgard MJ, Kozubal MA, Richardson TH *et al.*(2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* **5:** e9773.

Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB, YNP Metagenome Project Steering Committee *et al.*(2013). The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. *Front Microbiol* **4:** 67.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ *et al.*(2013). Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio* **4:** e00708-00713.

Konstantinidis KT, Tiedje JM (2005a). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102:** 2567-2572.

Konstantinidis KT, Tiedje JM (2005b). Towards a Genome-Based Taxonomy for Prokaryotes. *J Bacteriol* **187:** 6258-6264.

Koumandou VL, Kossida S (2014). Evolution of the F_0F_1 ATP Synthase Complex in Light of the Patchy Distribution of Different Bioenergetic Pathways across Prokaryotes. *Plos Comput Biol* **10**.

Kozubal MA, Romine M, Jennings R, Jay ZJ, Tringe SG, Rusch DB *et al.*(2013). Geoarchaeota: a new candidate phylum in the Archaea from high-temperature acidic iron mats in Yellowstone National Park. *ISME J* **7:** 622-634.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y *et al.*(2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40:** D123-129.

Martens JH, Barg H, Warren MJ, Jahn D (2002). Microbial production of vitamin B_{12}. *Appl Microbiol Biot* **58:** 275-285.

Meyer-Dombard DR, Swingley W, Raymond J, Havig J, Shock EL, Summons RE (2011). Hydrothermal ecotones and streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. *Environ Microbiol* **13:** 2216-2231.

Mitchell KR (2009). Controls on microbial community structure in thermal environments; exploring Bacterial diversity and the relative influence of geochemistry and geography. Ph.D. thesis, University of New Mexico, Albuquerque.

Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV (2007). Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol* **5:** 892-899.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH *et al.*(1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399:** 323-329.

Nesbo CL, Nelson KE, Doolittle WF (2002). Suppressive Subtractive Hybridization Detects Extensive Genomic Diversity in *Thermotoga maritima*. *J Bacteriol* **184:** 4475-4488.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al.*(2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39:** 3204-3223.

Posada D, Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Price PB (2009). Microbial genesis, life and death in glacial ice. *Can J Microbiol* **55:** 1-11.

Reysenbach AL, Wickham GS, Pace NR (1994). Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. *Appl Environ Microbiol* **60:** 2113-2119.

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

Say RF, Fuchs G (2010). Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* **464:** 1077-1081.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W *et al.*(2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7:** 539.

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

Stetter KO (2006). Hyperthermophiles in the history of life. *Philos T R Soc B* **361:** 1837-1842; discussion 1842-1833.

Streit WR, Entcheva P (2003). Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Appl Microbio and Biot* **61:** 21-31.

Sutcliffe IC (2011). Cell envelope architecture in the Chloroflexi: a shifting frontline in a phylogenetic turf war. *Environ Microbiol* **13:** 279-282.

Suzuki M, Hirai T, Arai H, Ishii M, Igarashi Y (2006). Purification, Characterization, and Gene Cloning of Thermophilic Cytochrome cd_1 Nitrite Reductase from *Hydrogenobacter thermophilus* TK-6. *J Biosci Bioeng* **101:** 391-397.

Takacs-Vesbach C, Inskeep WP, Jay ZJ, Herrgard MJ, Rusch DB, Tringe SG *et al.*(2013). Metagenome Sequence Analysis of Filamentous Microbial Communities Obtained from Geochemically Distinct Geothermal Channels Reveals Specialization of Three Aquificales Lineages. *Front Microbiol* **4**.

Takai K, Horikoshi K (1999). Genetic Diversity of Archaea in Deep-Sea Hydrothermal Vent Environments. *Genetics* **152:** 1285-1297.

Takami H, Noguchi H, Takaki Y, Uchiyama I, Toyoda A, Nishi S *et al.*(2012). A Deeply Branching Thermophilic Bacterium with an Ancient Acetyl-CoA Pathway Dominates a Subsurface Ecosystem. *PLoS One* **7:** e30559.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30:** 2725-2729.

Team RC (2014). R: A language and environment for statistical computing, 3.1.0 edn. R Foundation for Statistical Computing: Vienna, Austria.

Wu D, Jospin G, Eisen JA (2013). Systematic Identification of Gene Families for use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS One* **8:** e77033.

Wu M, Chatterji S, Eisen JA (2012). Accounting for Alignment Uncertainty in Phylogenomics. *PLoS One* **7:** e30288.

Yarza P, Yilmaz P, Pruesse E, Glockner FO, Ludwig W, Schleifer KH *et al.*(2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12:** 635-645.

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT *et al.*(2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci U S A* **106:** 5865-5870.

Tables

Table 4.1. Genome assembly statistics for 'Pyropristinus' Type 1 and Type 2, and Calescamantes populations from either Octopus and/or Bechler Springs, Yellowstone National Park.

Population	Source	ID	Size[a]	C[b]	G+C (%)	Contigs	# Genes	% Protein coding	Longest Contig[c]
'Pyropristinus' Type 1	Octopus	T1.1	1.02	47.0	44.3	116	1249	97.4	34.4
'Pyropristinus' Type 1	Bechler	T1.2	1.24	64.9	44.3	72	1464	97.0	59.2
'Pyropristinus' Type 2	Octopus	T2.1	1.1	72.2	28.9	117	1376	98.0	32.5
Calescamantes -OS	Octopus	T3.1	1.29	62.9	35.2	164	1569	98.6	26.1

[a] in Mbp
[b] Completeness: Estimated based on the average of three estimation methods: 1) tRNA aaRS complement, 2) 40 conserved universal prokaryotic housekeeping genes; Wu *et al.* 2013, 3) 178 conserved universal bacterial housekeeping genes; Martin *et al.* 2006
[c] in Kbp

Figures

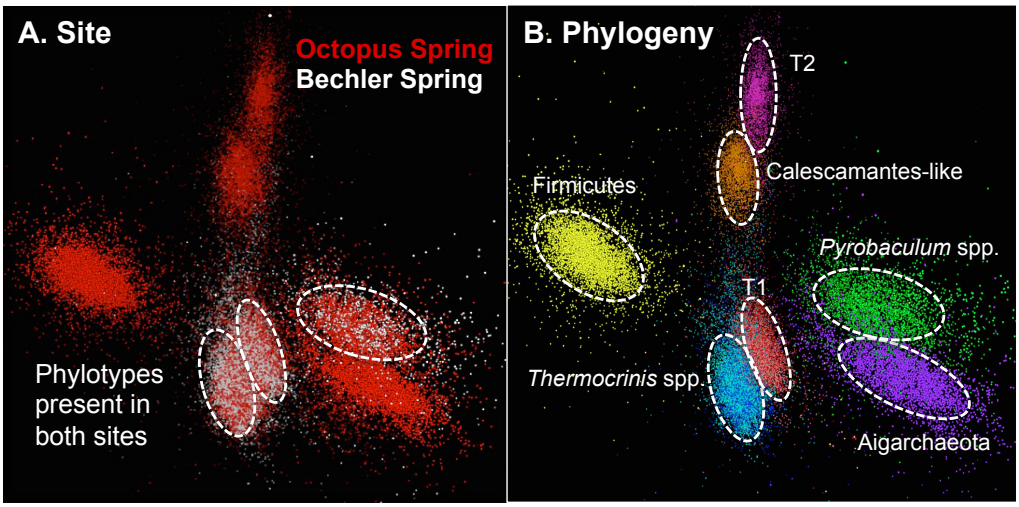


Figure 4.1. Nucleotide word frequency PCA plots of metagenome assemblies from two Aquificales 'streamer' communities in YNP. (A) Data colored by site: Octopus Spring = red; Bechler Spring = white. **(B)** Identical PCA orientation with phylogenetic analysis and assignment (dashed-white circles): 'Pyropristinus' Type 1-r01 = red; 'Pyropristinus' Type 1-r02 = light-red; 'Pyropristinus' Type 2-r01 = pink; Calescamantes-like = orange; Firm_T1-r01 = yellow; *Thermocrinis*-r01 = dark-blue; *Thermocrinis*-r02 = light-blue; *Pyrobaculum* spp. = green; Aigarchaeota_T1-r01 = purple.

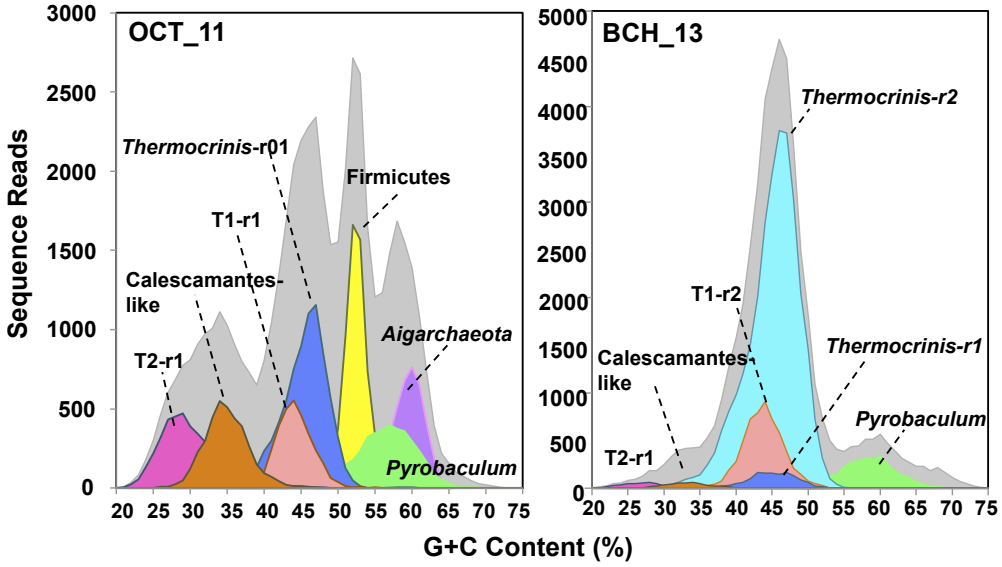


Figure 4.2. Frequency plots of the G+C content (%) of random shotgun sequence reads (Sanger) from filamentous 'streamer' communities at Octopus Spring (OCT_11) and Bechler Springs (BCH_13). Taxonomic (phylogenetic) assignment of each sequence read was performed using blastn (>90 % nt ID) against curated *de novo* assemblies generated from these sites (i.e., Figure 4.1): (light-gray = total reads, red = 'Pyropristinus' T1-r1 (G+C = 44 %), light-red = 'Pyropristinus' T1-r02 (G+C = 44 %), pink = 'Pyropristinus' T2-r01 (G+C = 29 %), orange = Calescamantes-like (G+C = 35 %), blue = *Thermocrinis*-like r01 (G+C = 45.5 %), light-blue = *Thermocrinis*-like r02 (G+C = 45 %), yellow = *Firmicutes* (G+C = 53 %), green = *Pyrobaculum*-like (G+C = 57-58 %), purple = Aigarchaeota (G+C = 60 %)).

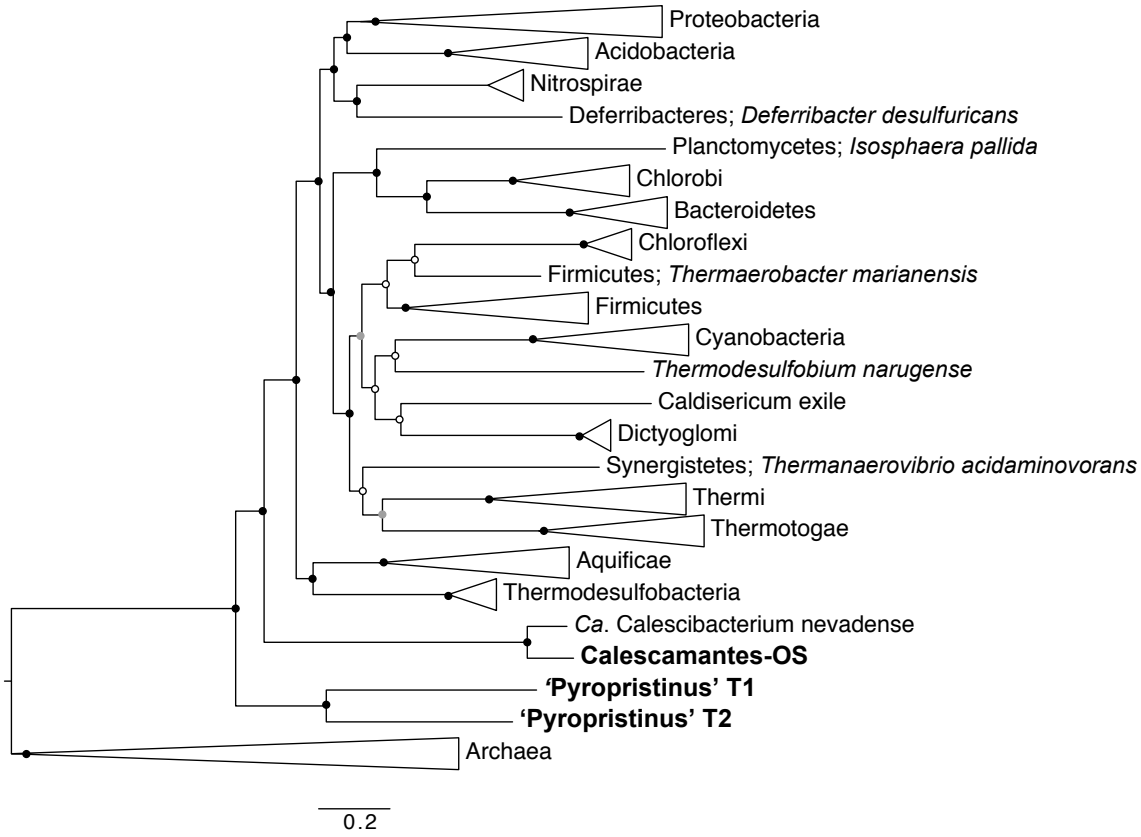


Figure 4.3. Phylogenomic analysis of 'Pyropristinus' and Calescamantes lineages. Maximum-likelihood tree based on genomic analysis incorporating 13 bacterial-specific and 5 universal housekeeping genes. Twenty-seven archaeal references were used as the outgroup. Phyla with more than one reference were collapsed. Bootstrap values are given at the nodes by solid black circles (90-100%), solid grey circles (50-89%) and open circles (< 50%). Scale shows expected substitutions per site.

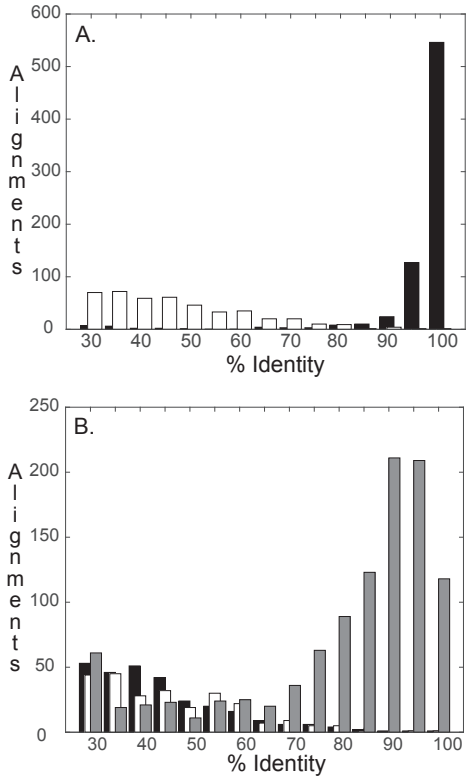


Figure 4.4. Distribution plots of amino acid identity % (AAI) of protein-coding genes between pairwise comparisons of novel bacterial assemblies. (**A)** The distribution of 'Pyropristinus' Type 1 from Octopus Spring (T1.1) vs. 'Pyropristinus' Type 1 from Bechler Spring (T1.2) (black; mean = 94.2% $\pm$ 10.1%), and 'Pyropristinus' Type 2 (T2.1 vs. T1.2) (white; mean = 46.6 $\pm$ 12.3%). (**B)** The distribution of the Calescamantes-like population from Octopus Spring vs. *Ca.* Calescibacterium nevadense (grey; mean = 78.0 $\pm$ 18.1%), 'Pyropristinus' Type 1 (T1.2) (black; mean = 42.7 $\pm$ 10.0%), and 'Pyropristinus' Type 2 (T2.1) (white; mean = 41.8 $\pm$ 9.0%).

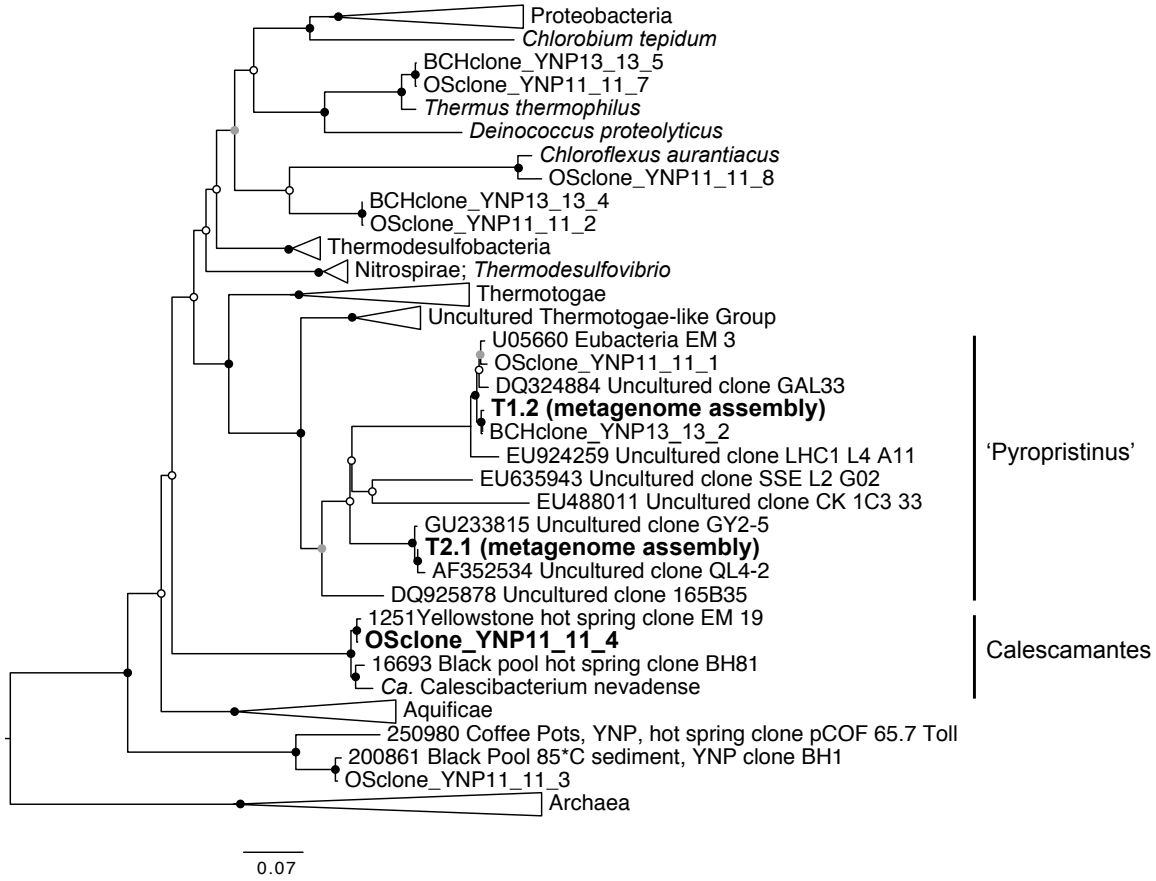


Figure 4.5. Phylogenetic analysis using near full-length 16S rRNA genes. 16S rRNA genes from the 'Pyropristinus' Types T1 and T2 assemblies are indicated in bold (T1.1 and Calescamantes-OS assemblies did not contain 16S rRNA genes and were thus omitted from this analysis). OSClone_YNP11_11_4, produced from a 16S rRNA gene library of the same Octopus Spring sample (also in bold) is nearly identical to the Calescamantes population from OS. Groups with multiple entries are collapsed as triangles. Bootstrap values of the maximum-likelihood tree are given at the nodes by solid black circles (90-100%), solid grey circles (50-89%) and open circles (< 50%).

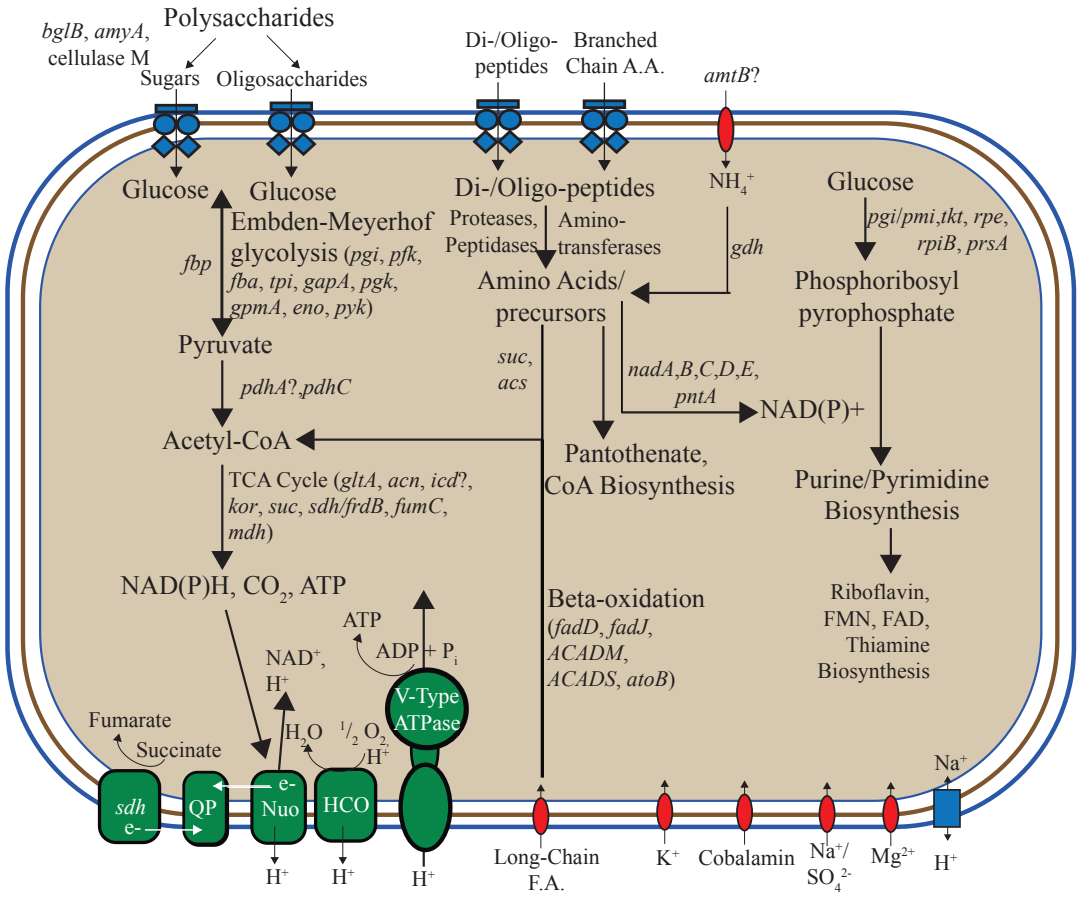


Figure 4.6. Metabolic reconstruction based on annotation and manual curation of 'Pyropristinus' Type 1 *de novo* assemblies. ABC-type transporters are represented as blue multi-component transmembrane proteins, other transporters as red transmembrane proteins, and the single antiporter as a blue rectangle. Complexes and enzymes used in aerobic respiration/ATP synthesis are identified in green. Gene abbreviations : *bglB*, phospho-β-glucosidase; *amyA*, α-amylase; LPS, Lipopolysaccharides; *nadA*, quinolinate synthase, *nadB*, L-aspartate oxidase; *nadC*, quinolinate phosphoribosyltransferase; *nadD*, nicotinate-mononucleotide adenylyltransferase; *nadE*, NAD synthetase; *pntA*, pyridine nucleotide transhydrogenase (α subunit); *pgi*, phosphoglucose isomerase; *pfk*, 6-

phosphofructokinase I; *fba*/*p*, fructose bisphosphate aldolase/phosphatase; *tpi*, triose phosphate isomerase; *gapA,* glyceraldehyde 3-phosphate dehydrogenase-A complex; *pgk*, phosphoglycerate kinase; *gpmA*, 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase; *eno*, enolase; *pyk*, pyruvate kinase; *pdhA*, pyruvate dehydrogenase (lipoamide); *pdhC*, pyruvate dehydrogenase E2 component; *gltA*, citrate synthase; *acn*, aconitate hydratase; *icd*, isocitrate dehydrogenase; *kor*, 2-oxoglutarate:ferredoxin oxidoreductase; *suc*, succinyl-CoA synthetase, *sdh*, succinate dehydrogenase; *frdB*, fumarate reductase iron-sulfur protein; *fumC*, fumarase C; *mdh*, malate dehydrogenase; *amtB*, ammonium transporter; *gdh*, glutamate dehydrogenase; *acs*, acetyl-CoA synthetase; *fadD*, fatty acyl-CoA synthetase; *fadJ*, 3-hydroxyacyl-CoA dehydrogenase; *ACADM*, acyl-CoA dehydrogenase (C-4 to C-12); *ACADS*, acyl-CoA dehydrogenase (C-2 to C-3); *atoB*, acetyl-CoA acetyltransferase; QP, quinone pool; *nuo*, NADH:ubiquinone oxidoreductose complex; HCO, Heme-Cu oxidase; *pgi*/*pmi*, glucose/mannose-6-phosphate isomerase; *tkt*, transketolase; *rpe*, ribulose-5-phosphate 3-epimerase; *rpiB*, ribose-5-phosphate isomerase B; *prsA*, ribose-phosphate pyrophosphokinase; FMN, flavin mononucleotide; FAD, flavin adenine dinucleotide. Question marks indicate genes not identified in either Type 1 assembly (list of identified protein-coding genes in Appendix File 4.A4).

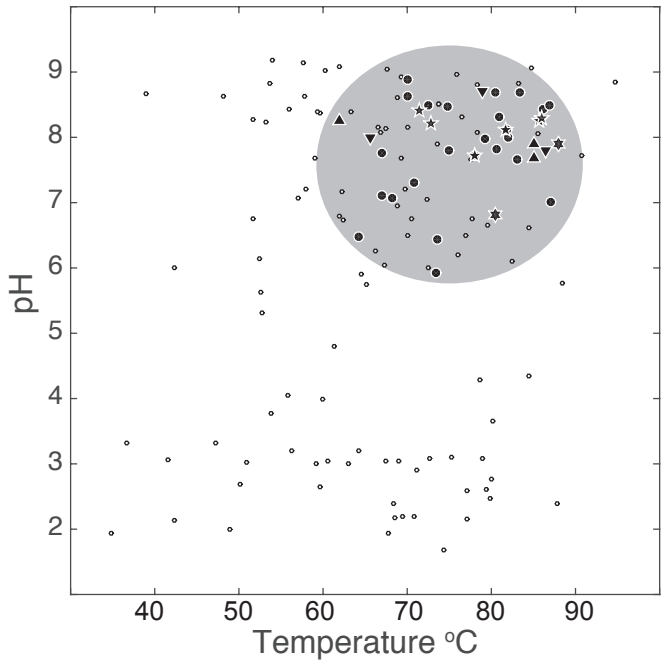


Figure 4.7. Temperature and pH distribution of the 'Pyropristinus' and Calescamantes phylotypes detected across different geothermal habitats. Sequences sharing > 97% nt identity to the 'Pyropristinus' and Calascamantes-OS population 16S rRNA genes were identified from prior 16S rRNA gene surveys of YNP and publically available databases [closed circles = 'Pyropristinus' Type 1 only (n = 24); upward facing triangles = 'Pyropristinus' Type 2 only (n = 3); downward facing triangles = Calescamantes-OS only (n = 3); stars = 2-3 out of three types present (n = 8); open circles = none of the three types]. Sites not containing these lineages are only shown for datasets that extensively surveyed YNP hot springs with universal bacterial PCR primers.

Supplementary Tables

Table 4.S1. Taxonomic information for references used in the phylogenomic analyses.

IMG ID	Domain	Phylum	Genome Name
639633064	Archaea	Crenarchaeota	Thermofilum pendens Hrk 5
640069332	Archaea	Crenarchaeota	Staphylothermus marinus F1, DSM 3639
640753029	Archaea	Crenarchaeota	Ignicoccus hospitalis KIN4/I, DSM 18386
641522657	Archaea	Crenarchaeota	Thermoproteus neutrophilus V24Sta
643348540	Archaea	Crenarchaeota	Desulfurococcus kamchatkensis 1221n
648028003	Archaea	Crenarchaeota	Acidilobus saccharovorans 345-15
648028062	Archaea	Crenarchaeota	Vulcanisaeta distributa DSM 14429
2510065009	Archaea	Crenarchaeota	Caldisphaera lagunensis IC-154, DSM 15908
637000163	Archaea	Euryarchaeota	Methanosphaera stadtmanae DSM 3091
638154502	Archaea	Euryarchaeota	Archaeoglobus fulgidus VC-16, DSM 4304
638154505	Archaea	Euryarchaeota	Methanocaldococcus jannaschii DSM 2661
638154509	Archaea	Euryarchaeota	Methanosarcina mazei Go1, DSM 3647
638154514	Archaea	Euryarchaeota	Pyrococcus abyssi GE5
638154521	Archaea	Euryarchaeota	Thermoplasma acidophilum DSM 1728
644736411	Archaea	Euryarchaeota	Thermococcus gammatolerans EJ3
646564501	Archaea	Euryarchaeota	Aciduliprofundum boonei T469
2511231109	Archaea	Euryarchaeota	Methanococcus maripaludis X1
2513237398	Archaea	Euryarchaeota	Thermococcus litoralis DSM 5473
2517093039	Archaea	Euryarchaeota	Pyrococcus furiosus COM1
2529293245	Archaea	Euryarchaeota	Haloferax mediterranei ATCC 33500
2531839487	Archaea	Euryarchaeota	Halosimplex carlsbadense 2-9-1
2554235477	Archaea	Euryarchaeota	Natronococcus amylolyticus DSM 10524
2504756013	Archaea	Geoarchaeota	Geoarchaeota archaeon OSPB-1 (8 best

			scaffolds)
641522611	Archaea	Korarchaeota	Candidatus Korarchaeum cryptofilum OPF8
641228499	Archaea	Thaumarchaeota	Nitrosopumilus maritimus SCM1
651324018	Archaea	Thaumarchaeota	Candidatus Nitrosoarchaeum limnia SFB1
2510065023	Archaea	Thaumarchaeota	Candidatus Nitrososphaera gargensis Ga9-2
643692001	Bacteria	Acidobacteria	Acidobacterium capsulatum ATCC 51196
639633060	Bacteria	Acidobacteria	Candidatus Solibacter usitatus Ellin6076
644736401	Bacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. trifolii WSM1325
637000010	Bacteria	Aquificae	Aquifex aeolicus VF5
650377953	Bacteria	Aquificae	Hydrogenobacter thermophilus TK-6, DSM 6534
2506210035	Bacteria	Aquificae	Hydrogenobaculum sp. 3684 ((Finished QAed))
2506210034	Bacteria	Aquificae	Hydrogenobaculum sp. HO
2507262044	Bacteria	Aquificae	Hydrogenobaculum sp. SN
642555132	Bacteria	Aquificae	Hydrogenobaculum sp. Y04AAS1
643692050	Bacteria	Aquificae	Sulfurihydrogenibium azorense Az-Fu1
642555165	Bacteria	Aquificae	Sulfurihydrogenibium sp. YO3AOP1
646564582	Bacteria	Aquificae	Thermocrinis albus HI 11/12, DSM 14484
2512875013	Bacteria	Aquificae	Thermocrinis ruber DSM 12173
2517572120	Bacteria	Bacteroidetes	Bacteroides barnesiae DSM 18169
637000025	Bacteria	Bacteroidetes	Bacteroides fragilis YCH46
642555148	Bacteria	Bacteroidetes	Porphyromonas gingivalis ATCC 33277
648028028	Bacteria	Betaproteobacteria	Gallionella capsiferriformans ES-2
637000195	Bacteria	Betaproteobacteria	Nitrosomonas europaea ATCC 19718
637000197	Bacteria	Betaproteobacteria	Nitrosospira multiformis ATCC 25196
2513237181	Bacteria	Caldiserica	Caldisericum exile AZM16c01, NBRC 104410

2527291514	Bacteria	Calescamantes	Calescamantes bacterium JGI 0000106-G12 (
642555121	Bacteria	Chlorobi	Chlorobium limicola DSM 245
637000073	Bacteria	Chlorobi	Chlorobium tepidum TLS
642555123	Bacteria	Chlorobi	Chloroherpeton thalassium ATCC 35110
643348527	Bacteria	Chloroflexi	Chloroflexus aggregans DSM 9485
641228485	Bacteria	Chloroflexi	Chloroflexus aurantiacus J-10-fl
640069323	Bacteria	Cyanobacteria	Prochlorococcus marinus MIT 9303
2506520048	Bacteria	Cyanobacteria	Synechococcus sp. PCC 7336
646564525	Bacteria	Deferribacteres	Deferribacter desulfuricans SSM1, DSM 14783
2503692001	Bacteria	Deltaproteobacteria	Desulfococcus oleovorans Hxd3
643348539	Bacteria	Deltaproteobacteria	Desulfovibrio vulgaris Miyazaki F
639633022	Bacteria	Deltaproteobacteria	Desulfovibrio vulgaris vulgaris DP4
643348542	Bacteria	Dictyoglomi	Dictyoglomus thermophilum H-6-12, ATCC 35947
643348543	Bacteria	Dictyoglomi	Dictyoglomus turgidum DSM 6724
646311904	Bacteria	Firmicutes	Ammonifex degensii KC4
649633022	Bacteria	Firmicutes	Caldicellulosiruptor hydrothermalis 108
637000060	Bacteria	Firmicutes	Carboxydothermus hydrogenoformans Z-2901
640069310	Bacteria	Firmicutes	Desulfotomaculum reducens MI-1
649633101	Bacteria	Firmicutes	Thermaerobacter marianensis 7p75a, DSM 12885
2503538027	Bacteria	Firmicutes	Thermoanaerobacter ethanolicus JW 200
2504756006	Bacteria	Firmicutes?	Thermodesulfobium narugense Na82, DSM 14796
2529292882	Bacteria	Gammaproteobacteria	Escherichia coli B799
637000194	Bacteria	Gammaproteobacteria	Nitrosococcus oceani C-107, ATCC 19707
2505313057	Bacteria	Gammaproteobacteria	Pseudomonas syringae Pph1302A (Psy97)
640427151	Bacteria	Gammaproteobacteria	Vibrio cholerae O395

2529292556	Bacteria	Nitrospirae	Thermodesulfovibrio islandicus DSM 12570
643348581	Bacteria	Nitrospirae	Thermodesulfovibrio yellowstonii DSM 11347
2523533630	Bacteria	Nitrospirae	Thermodesulfovibrio thiophilus DSM 17215
649633058	Bacteria	Planctomycetes	Isosphaera pallida IS1B, ATCC 43644
646311961	Bacteria	Synergistetes	Thermanaerovibrio acidaminovorans Su883
649633035	Bacteria	Thermi	Deinococcus proteolyticus MRP, DSM 20540
2505679077	Bacteria	Thermi	Thermus thermophilus SG0.5JP17-16
2506520012	Bacteria	Thermodesulfobacteria	Thermodesulfobacterium geofontis OPF15
2524023142	Bacteria	Thermodesulfobacteria	Thermodesulfobacterium hveragerdense DSM 12571
2523533618	Bacteria	Thermodesulfobacteria	Thermodesulfobacterium thermophilum DSM 1276
640753026	Bacteria	Thermotogae	Fervidobacterium nodosum Rt17-B1
2507149014	Bacteria	Thermotogae	Fervidobacterium pennivorans Ven 5, DSM 9078
643348583	Bacteria	Thermotogae	Thermosipho africanus TCF52B
640753057	Bacteria	Thermotogae	Thermosipho melanesiensis BI429
641228511	Bacteria	Thermotogae	Thermotoga lettingae TMO
2519899531	Bacteria	Thermotogae	Thermotoga maritima MSB8, DSM 3109
646311964	Bacteria	Thermotogae	Thermotoga naphthophila RKU-10
643348584	Bacteria	Thermotogae	Thermotoga neapolitana DSM 4359
640427150	Bacteria	Thermotogae	Thermotoga petrophila RKU-1
2531839610	Bacteria	Thermotogae	Thermotoga sp. EMP
642487181	Bacteria	Thermotogae	Thermotoga sp. RQ2
2503508007	Bacteria	Thermotogae	Thermotoga thermarum LA3, DSM 5069

Table 4.S2. Abundance estimates of major population types based on analysis of metagenome sequence reads against reference *de novo* assemblies (90% nucleotide identity) obtained from Octopus and Bechler Springs, Yellowstone National Park (populations shaded are the focus of the current study).

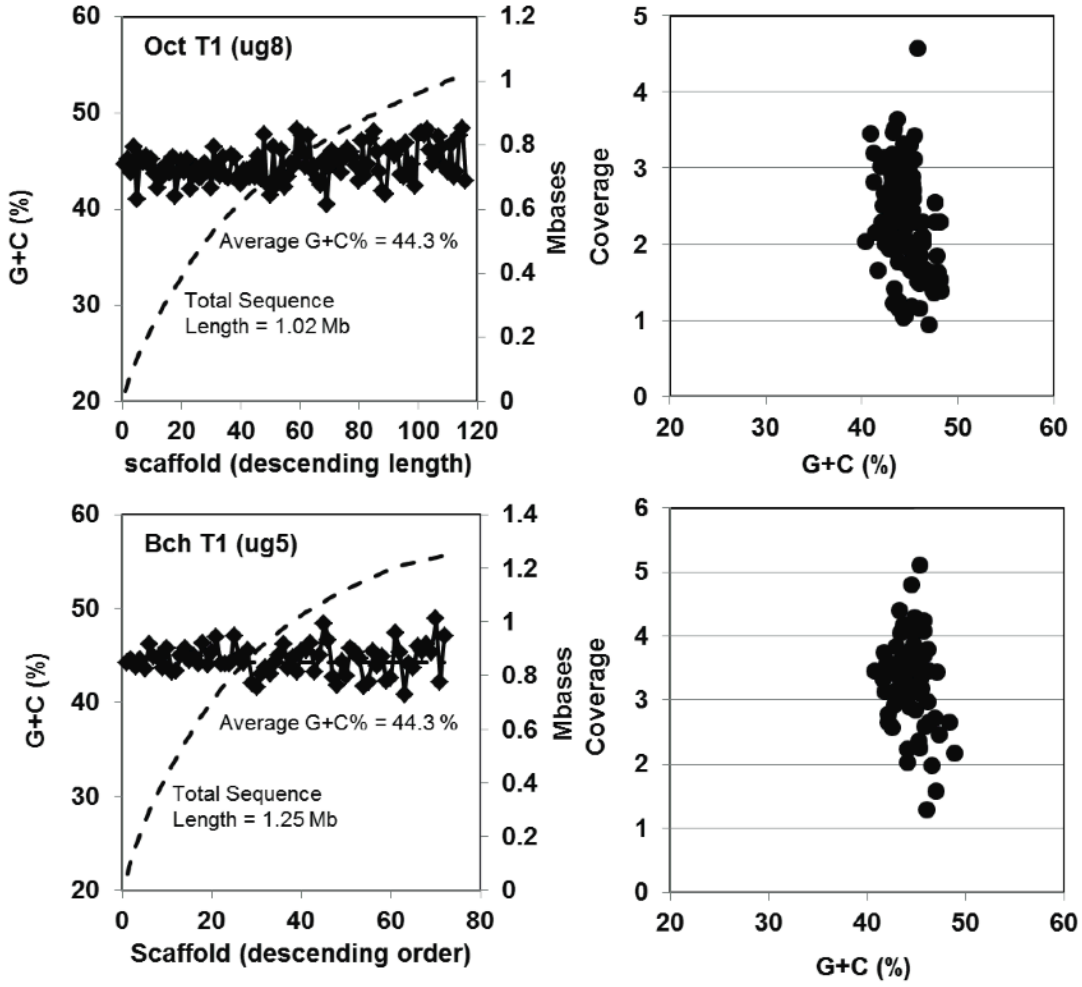
Phylotype	G+C (%)	Octopus Spring	Bechler Spring
Aquificales, *Thermocrinis* spp.	45	18	62
Novel Firmicutes	53	13	< 1
'Pyropristinus', Type 1	44	8	12
'Pyropristinus', Type 2	29	8	< 1
Aigarchaeota, Ca. *Calditenius rheumensis*[a]	60	8	< 1
Pyrobaculum spp.	57	7	7
Calescamantes-OS[b]	35	7	< 1
Unknown Bacteria		3	< 1
Desulfurococcales	58	< 1	< 1
Orphans		25	13

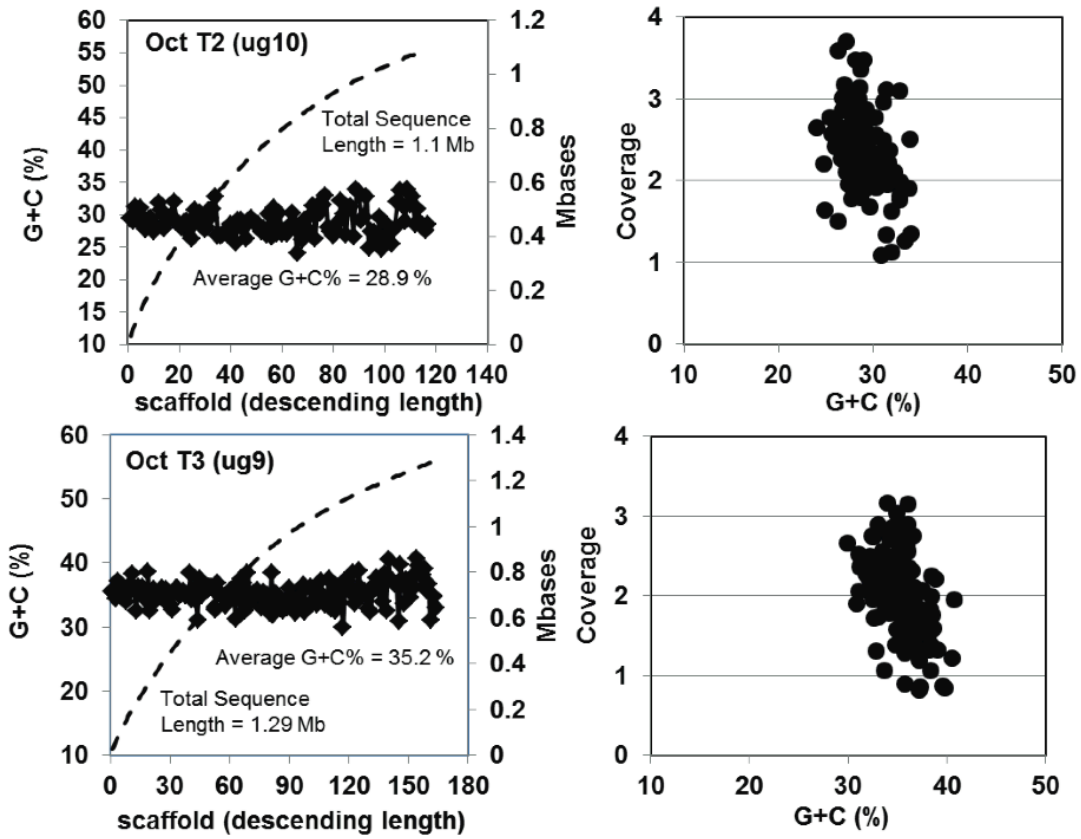
[a] Beam et al., 2015 (submitted)

[b] Hedlund et al., 2014

Supplementary Figures

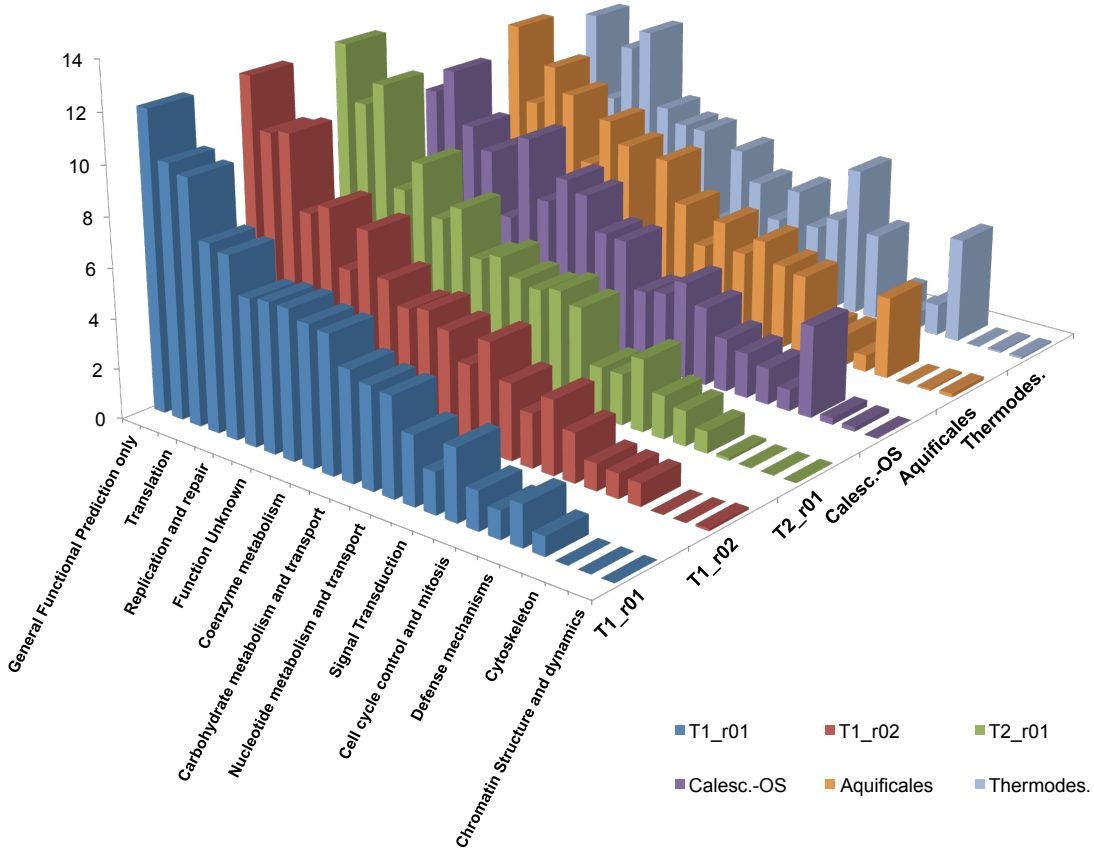
Figure 4.S1. Coverage and G + C (%) analysis of contigs corresponding to the four assembled populations.





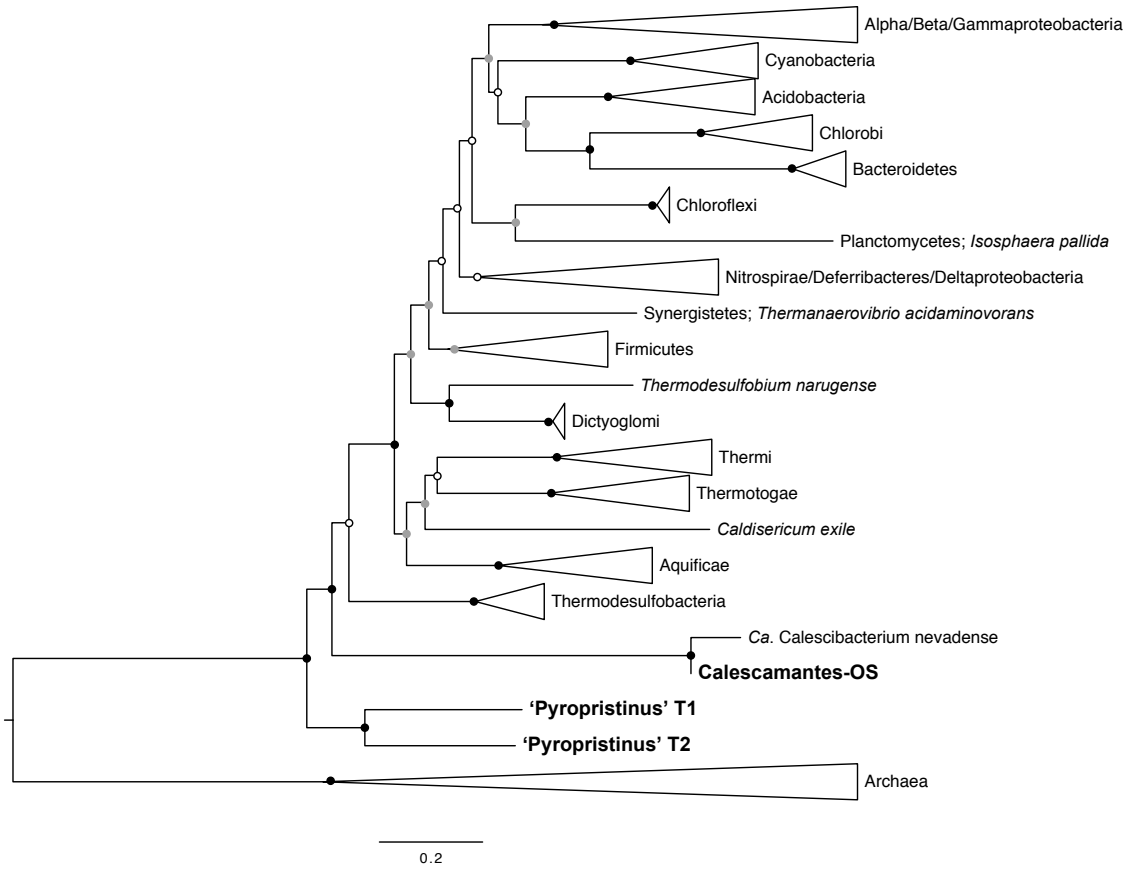
Cumulative sequence is reported on the right y-axis (dashed line) on the left panels and individual scaffolds are plotted by G+C % (left y-axis). Right panels show G+C (%) as a function of scaffold coverage for (**A)** T1.1 (top) and T1.2 (bottom) and **(B)** T2.1 (top) and T3.1 (bottom)

Figure 4.S2. Relative gene frequency (%) of COG categories for all four assemblies and closely-related phyla.



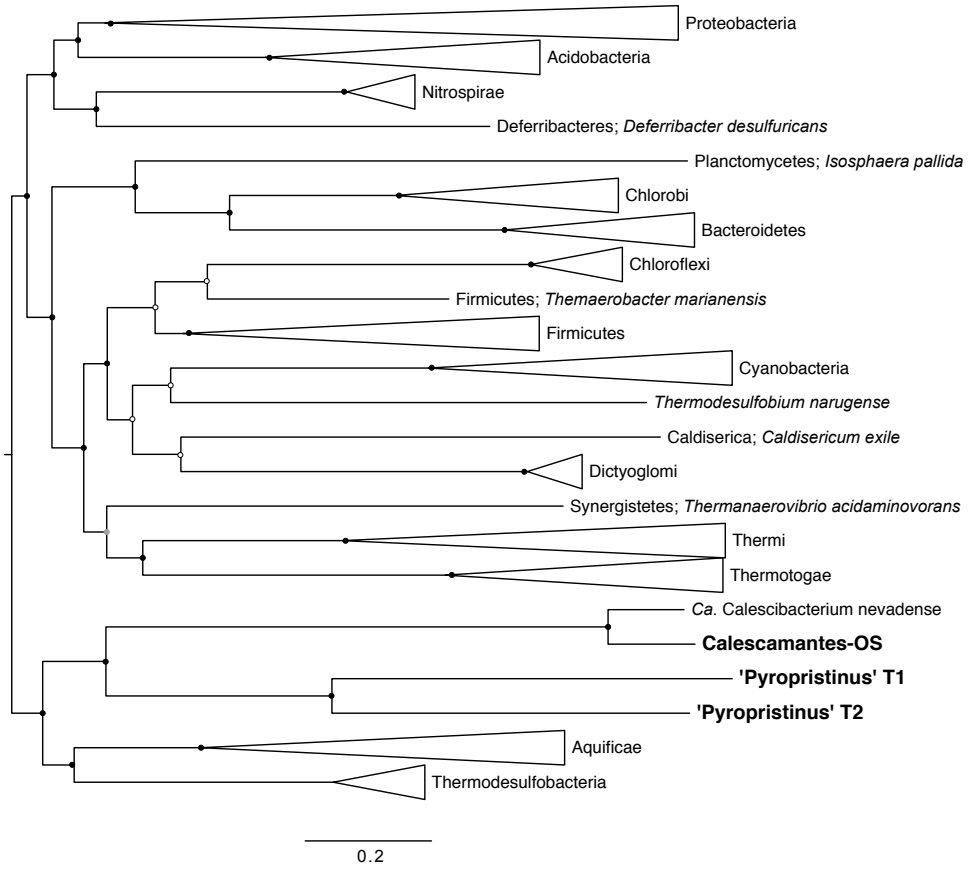
Relative abundance of each COG category is given on the Y axis for the 'Pyropristinus' T1 and T2, Calescamantes-OS and the related *Aquificales* and *Thermodesulfobacteria*.

Figure 4.S3. Maximum likelihood phylogenetic tree of ribosomal proteins of the 'Pyropristinus' and Calescamantes lineages.



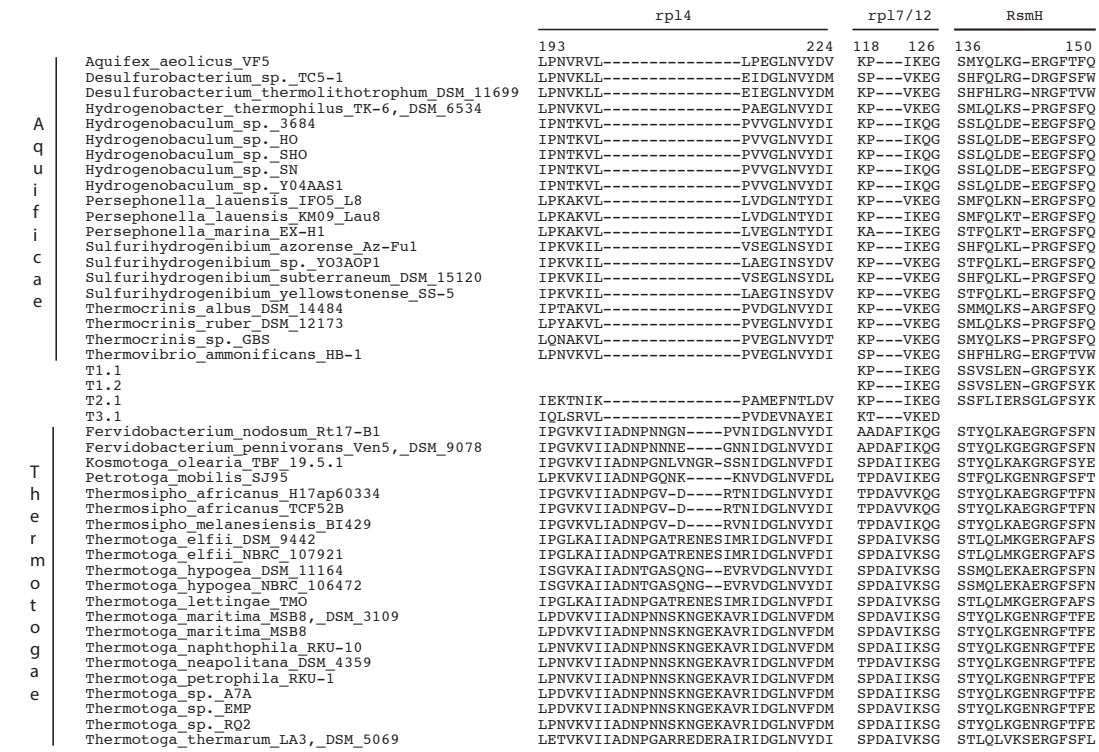
ML tree of a 5 (4 universal, 1 bacterial-only) ribosomal protein concatenation. Twenty-seven archaeal references were used as outgroups. Phyla with more than one reference are collapsed. Circles at nodes correspond to bootstrap percentages (out of 100 ML

Figure 4.S4. Maximum likelihood phylogenetic tree of the 'Pyropristinus' and Calescamantes lineages without outgroups.



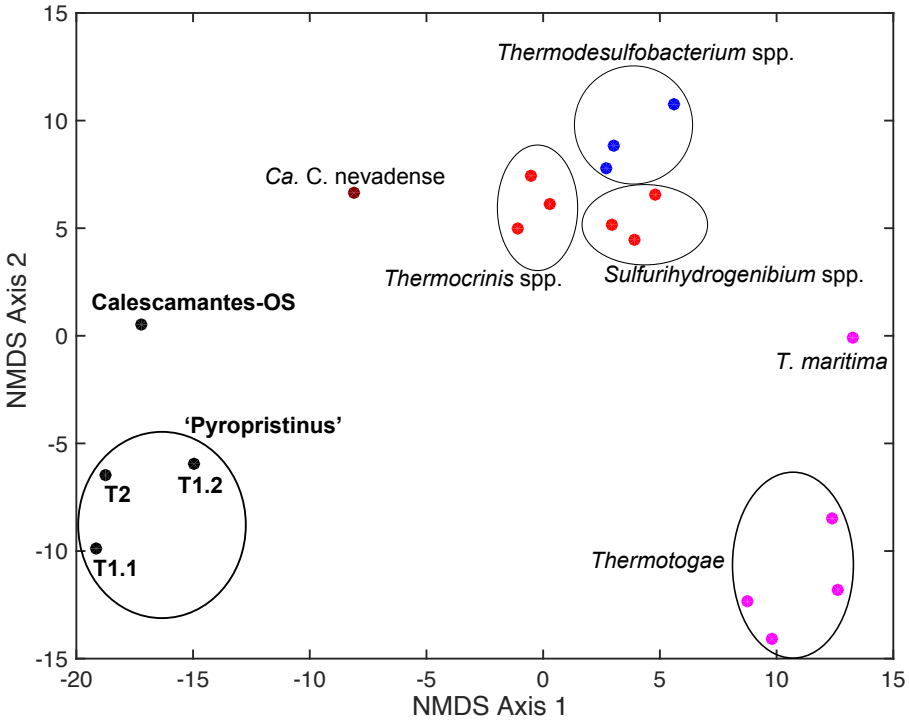
ML tree of a concatenated single-copy housekeeping gene dataset with 13 bacterial-specific and 5 universal housekeeping genes. Phyla with more than one reference are collapsed. Circles at nodes correspond to bootstrap percentages (out of 100 ML replicates): black circles (90-100%), grey circles (50-89%) and white circles (< 50%). Scale shows expected substitutions per site.

Figure 4.S5. Partial amino acid alignments for genes previously found to contain conserved phylum-specific indels for *Thermotogae* and *Aquificae*.



50S ribosomal protein L4 contains a *Thermotogae*-specific 10-15 amino acid (a.a.) insertion beginning at position 200. The 50S ribosomal protein L7/L12 contains a *Thermotogae*-specific 3 a.a. insertion beginning at position 120. Ribosomal RNA small subunit methyltransferase H (RsmH) contains an *Aquificae*-specific 1 a.a. deletion beginning at position 143. Genes not present in the assemblies are indicated by blank lines.

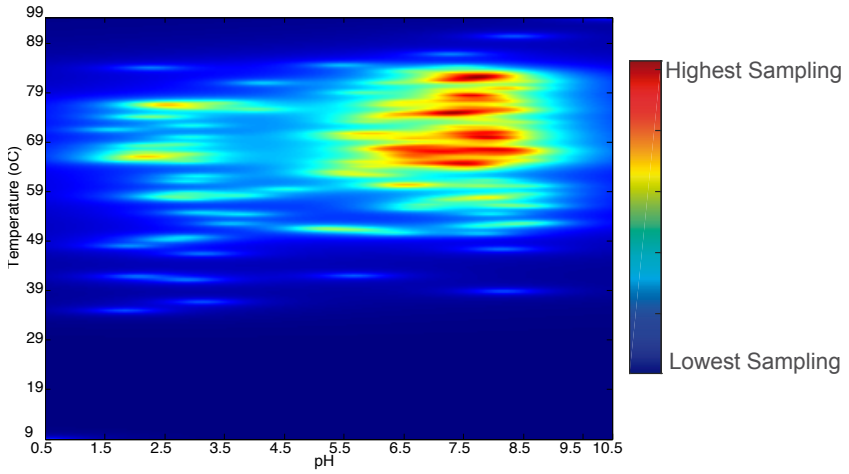
Figure 4.S6. Non-metric multidimensional scaling plots (NMDS) for COG distribution amongst 'Pyropristinus', Calescamantes and closely related lineages.



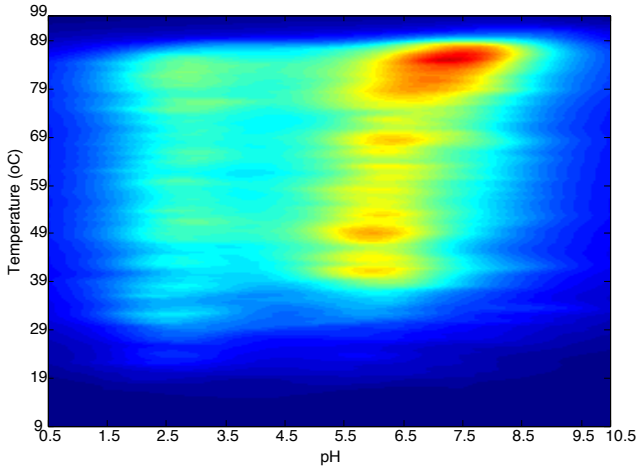
NMDS plots were constructed from presence/absence Euclidean-distance matrices of COG groups present in the 'Pyropristinus' T1 and T2 assemblies and Calescamantes-OS in addition to *Ca.* Calescibacterium nevadense (dark red), a subset of *Aquificae* (bright red), *Thermodesulfobacteria* (blue) and *Thermotogae* (purple) genomes, which comprised the closest related lineages to the 'Pyropristinus' and Calescamantes lineages (NMDS stress = 0.08).

Figure 4.S7. Sampling density maps for YNP geothermal ecosystem and sites used to infer 'Pyropristinus' T1 and T2 and Calescamantes-OS distribution.

A.



B.



(A) Sampling density of springs used to infer distribution of 'Pyropristinus' T1, T2 and Calescamantes-OS by temperature and pH of springs (n=141). **(B)** Sampling density of springs in the YNP Research Coordination Network database (n=7680) by temperature and pH. Color scale indicates normalized sampling density for each dataset with dark blue indicating lowest density and dark red indicating highest density.

Chapter 5

Conclusions

The studies described here contribute to our understanding of thermal spring biodiversity by expanding our understanding of the archaeal phylogenetic diversity present in Yellowstone National Park (YNP) springs and assessing the metabolic potential of ubiquitous, uncultured bacterial populations.

In chapter 2, I demonstrated that there exists a methodological and systematic bias in the detection of uncultured archaeal populations by use of inadequate 16S rRNA gene PCR primers. In using a multifaceted approach to describe the archaeal community composition of a high temperature circumneutral spring in YNP, I identified archaeal 16S rRNA gene PCR primers that would accurately capture the most archaeal phylogenetic diversity present in thermal spring microbial consortia. The results indicated that traditional, commonly used PCR primers were inadequate to capture the phylogenetic diversity, but that less commonly used primers could provide more accurate results.

Using the 16S rRNA gene PCR primers identified in Chapter 2, I surveyed the phylogenetic diversity of Archaea in the YNP thermal spring ecosystem using high-throughput 16S rRNA gene 454 pyrosequencing and documented a surprisingly diverse and ubiquitous distribution of archaeal phylogenetic diversity in YNP. This is the largest study to date (in sequencing depth and in sample number) to survey archaeal diversity across YNP, and the only study to integrate both bacterial and archaeal high-throughput data to contrast diversity patterns between the two domains across the YNP system. The results reported here largely suggest that the diversity of the two domains is structured

across YNP thermal springs similarly. Some notable differences were observed in diversity partitioning (e.g. differing relationships to temperature and [SO_4^{2-}] as an indicator of higher archaeal diversity), and further analysis of these parameters may provide fundamental insight into ecological differences between the two domains. The results discussed here also provide a rigorous statistical analysis of the major abiotic parameters associated with archaeal and bacterial community composition. Our results are concordant with the model based on metagenomic community differences proposed by Inskeep *et al.* 2013 where pH, followed by temperature and then geochemical components (e.g. S^{2-} and O_2) structure communities hierarchically. My results are also consistent with several other recent reports that pH and temperature structure thermal spring communities globally. The results described here also provide a relative measure of the effects of these parameters, which is often lacking due to undersampling across large ranges of temperature and pH. Lastly, the co-occurrence results discussed here suggest population targets for *in-situ* community function analyses (e.g. the Aigarchaeota) that may not have been previously recognized as important components of thermal spring communities due to the lack of detection and datasets that do not permit across-domain cooccurrence analyses.

In Chapter 4, I described the genomic characterization of multiple, deep-branching and novel bacterial lineages. The vast amount of microbial phylogenetic diversity that is without cultured representatives or genomic references to infer physiologic potential has spurred major initiatives to fill in these taxonomic gaps and the analyses reported here contributes to these efforts. From an evolutionary perspective, the "Pyropristinus" bacterial populations are integral to understanding the evolution of the

earliest bacterial lineages because of their phylogenetic placement near the root of all Bacteria. Physiological differences between the proposed "Pyropristinus" division and other, closely related Bacteria provide opportunities to understand the nature of divergence in metabolic strategies in the deepest-diverging bacterial lineages (e.g. divergence in ATP synthase variants and respiratory capabilities). From an ecological perspective, the characterization of the metabolic potential of these populations provides an excellent opportunity to probe community-wide interactions (e.g. heterotrophic-autotrophic interactions) within 'streamer' microbial communities, which now have most major populations genomically characterized (from the results reported here and others). Further, the characterization of these ubiquitously distributed, neutrophilic, hyperthermophiles provides a physiologic basis for understanding the ecological interactions underpinning their distribution and diversity across YNP springs.

List of Appendix Files

Appendix File 3.A1. Sample location and geochemical measurements for each spring.

Appendix File 4.A2. Metadata and sample information for samples from databases/datasets that were used to survey the distribution of the three lineages.

Appendix File 4.A3. List of phylum-specific CSIs for Thermotogae and Aquificae that were used as references for the T1, T2 and T3 lineages

Appendix File 4.A4. List of pathways with presence/absence of genes used to infer the metabolic potential of the lineages.