Winter 2020

Genomic Insights and Ecological Adaptations of Deep-Subsurface and Near Subsurface Thermococcus Isolates

Lilja Caitlin Strang
Western Washington University, trustedsamurai7000@gmail.com

Genomic Insights and Ecological Adaptations of Deep-Subsurface and Near Subsurface *Thermococcus* Isolates

By

Lilja C. Strang

Accepted in Partial Completion
of the Requirements for the Degree
Master of Science

ADVISORY COMMITTEE

Dr. Craig Moyer, Chair

Dr. Dietmar Schwarz

Dr. Jeff Young

Dr. Heather Fullerton

GRADUATE SCHOOL

David L. Patrick, Interim Dean

Master's Thesis

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Western Washington University, I grant to Western Washington University the non-exclusive royalty-free right to archive, reproduce, distribute, and display the thesis in any and all forms, including electronic format, via any digital library mechanisms maintained by WWU.

I represent and warrant this is my original work, and does not infringe or violate any rights of others. I warrant that I have obtained written permissions from the owner of any third party copyrighted material included in these files.

I acknowledge that I retain ownership rights to the copyright of this work, including but not limited to the right to use all or part of this work in future works, such as articles or books.

Library users are granted permission for individual, research and non-commercial reproduction of this work for educational purposes only. Any further digital posting of this document requires specific permission from the author.

Any copying or publication of this thesis for commercial purposes, or for financial gain, is not allowed without my written permission.

Lilja Strang

February 28th, 2020

Genomic Insights and Ecological Adaptations of Deep-Subsurface and Near Subsurface *Thermococcus* Isolates

A Thesis
Presented to
The Faculty of
Western Washington University

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

by
Lilja Strang
February 2020

ABSTRACT

Members of the Archaeal genus *Thermococcus* are sulfur-dependent hyperthermophiles found in hydrothermal vents throughout the world. Previous analysis of a *Thermococcus* culture collection containing isolates from the Juan de Fuca Ridge, Gorda Ridge, and South East Pacific Rise using amplified fragment length polymorphism analysis and multilocus sequence typing revealed a distinct clade of *Thermococcus* isolated from the 1996 megaplume event at Gorda Ridge, indicating that they originated from a deep-subsurface habitat. The aim of this study was to elucidate the functional adaptations that allow for the survival of the Gorda Ridge clade in a deep-subsurface habitat as compared to representative *Thermococcus* isolates from shallow subsurface environments. This was accomplished through a pangenomic analysis of representative isolates in this clade and others from this culture collection. The Gorda Ridge megaplume group was enriched for genes relating to DNA repair and stabilization including a predicted endonuclease distantly related to Archaeal Holliday junction resolvase, DNA mismatch repair ATPase *mutS*, CRISPR/Cas elements, and *dnaK* (*hsp70*). The group was also enriched for ABC-type branched-chain amino acid (BCAA) transport system, enzymes for the Shikimate pathway for aromatic amino acid synthesis, as well as TupA for tungstate transport. These findings suggest that *Thermococcus* inhabiting deep-subsurface fluid reservoir require the added ability to prevent and repair damage to their DNA, presumably due to the energy demands of DNA replication. The enrichment in BCAA and tungstate transporters may indicate the use of an amino acid catabolism pathway followed by fermentation catalyzed by the tungstopterin containing enzymes aldehyde ferredoxin oxidoreductase and alcohol dehydrogenase, suggesting a preference for peptides over carbohydrates as an energy source in the deep-subsurface.

ACKNOWLEDGEMENTS

I would like to thank the Center for Dark Energy Biosphere Investigations for the award of the C-DEBI Graduate Student Fellowship, and the Western Washington University Enhancement of Graduate Research Fund for their financial support. I would like to thank WWU and my committee members. I would also like to thank the undergraduate research assistants from Craig Moyer's lab for their assistance in maintaining the *Thermococcus* culture collection: Kyle Neal, Nathan Wilks, Anna Ratzlaff, Daria Gausman, Davin Hoover, Jaykish Patel, and Lindsay Heimerl. Additionally, I would like to thank my family members Marie Strang, Timothy Strang, Sylvia Strapps-Coon, and Lavern Coon for their support while researching and writing my thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The classical view of microbial biogeography is that "everything is everywhere; environment selects" (Baas-Becking, 1934), meaning microbial populations exhibit high dispersal, but are shaped by environmental parameters into habitat specific ecotypes and species. However, not all microbial populations follow this pattern. For example, geographic barriers isolate regional populations of *Sulfolobus* leading to region specific differentiation that is independent of environmental parameters (Whitaker et al., 2003). Furthermore, other microorganisms such as the bacterioplankton *Pelagibacter ubique* are well known for having a global geographic range (Morris et al., 2002). Complicating matters further are hyperthermophiles such as *Thermococcus*. While these organisms exhibit biogeographic patterns of high dispersal, niche adaptations have led to incidences of habitat specific divergence and speciation (Price et al., 2015). Microorganisms do not adhere to simple models of broad dispersal or allopatric speciation (Whitaker, 2006). Therefore, in order to gain a more complete understanding of microbial evolution it is important to interrogate individual populations for unique biogeographic signals, such as functional adaptations that may indicate region or habitat specific ecotypes which could signify the first step to species differentiation.

Adaptations to an extreme environment have led to diverse microbial metabolisms and physiologies in hydrothermal vent systems (Mayer & Müller, 2014; Valentine, 2007). Deep-sea microorganisms are of interest not only because their collective biomass may be comparable to the biomass of all surface life (Gold, 1992; Whitman et al., 1998; McMahon & Parnell, 2013), but also because they play important roles in biogeochemical cycles by metabolizing nitrogen, sulfur, hydrogen, and hydrocarbon containing compounds into biologically accessible products (Lovely & Chapelle, 1995; Dick et al., 2013). Deep-sea microbes also have the potential to provide novel enzymes for biotechnology applications (Alma'abadi et al., 2015). For example, DNA polymerase

from *Pyrococcus furiosus*, a hyperthermophilic archaeon found in hydrothermal vents, was found to have higher fidelity than *Taq* DNA polymerase (Saiki et al., 1988) when used in PCR amplification due to its 3' to 5' proofreading exonuclease activity (Lundberg et al., 1991). Hyperthermophiles have also been studied for their use in the field of industrial bioenergy. *P. furiosus* and several strains of *Thermococcus* have the ability to produce large amounts of H_2 gas, which could be applied in the conversion of organic feedstock into H_2 for use as a fossil fuel alternative (Oslowski et al., 2011).

Microorganisms from the order *Thermococcales*, which includes the genera *Pyrococcus* (Fiala & Stetter, 1986), *Paleococcus* (Takai et al., 2000), and *Thermococcus* (Zillig et al., 1983). Members of the order *Thermococcales* are diverse and show evidence of unique adaptations such as the structural variations of ATP synthase in the genera *Pyrococcus* and *Thermococcus*, the use of formate oxidation for carbon and energy by *Thermococcus onnurineus* (Kim et al., 2010), and ferredoxin reduction by *Pyrococcus furiosus* for sugar oxidation (Sapra et al., 2003). Of these genera, *Thermococcus* is found in particularly large numbers in hydrothermal vent systems (Pledger & Baross, 1991). *Thermococcus spp.* are hyperthermophilic anaerobic heterotrophs that ferment organic compounds and occupy a variety of niches. Some strains use elemental sulfur (S^0) as an electron acceptor, resulting in the production of H_2S (Robb & Place, 1995; Teske et al., 2009) and making them an integral part of the sulfur cycle.

When vent fluids from the subseafloor are released during eruptions they bring with them microorganisms native to deep-subsurface habitats. These eruptions provide a rare opportunity to study organisms that originate at a depth normally only accessible through deep-sea drilling expeditions (Summit & Baross, 1998). One such seismic event happened in February 1996 in the North Gorda Ridge spreading center 300 km from the coast of Oregon and Southern California.

This event was associated with a dike intrusion that caused the sudden eruption of a large amount of hot hydrothermal vent fluid, also called a megaplume, rather than the slow leaking of fluid that is usually characteristic of these systems (Baker, 1998; Chadwick et al., 1998; Fox & Dziak, 1998). Megaplumes are a plausible mechanism of dispersal for microbes otherwise confined to the island-like ecosystem of deep-sea hydrothermal vents (Dobbs & Selph, 1997). The pre-eruption fluid resides in the reservoir for months or years at a time (Lupton, 1996), providing adequate opportunity for a deep-subsurface ecotype to emerge (Summit & Baross, 1998). *Thermococcus* residing in the fluid reservoir are carried up to the water column during the eruption event by the buoyant plume fluid (Lupton et al., 1999). *Thermococcus* originating from the deep-subsurface fluid reservoir and brought up to the surrounding water column via the megaplume were isolated through serial dilution to extinction (Summit & Baross, 1998). Traits have been observed in *Thermococcus* and *Pyrococcus* that allow them to survive in cold, oxygenated seawater for extended periods of time (Jannasch et al., 1992) which could help them survive long distance dispersal as the plume travels thousands of kilometers from its source (Lupton et al., 1998).

Ecological and metabolic diversity within *Thermococcus* has been observed in isolates from different habitats within the same vent site. *Thermococcus* isolates from the 1996 megaplume event were analyzed using amplification and sequencing of the small subunit rRNA gene and small and large subunit intergenic spacer region, showing that subseafloor isolates were phylogenetically distinct from their counterparts from sulfide chimneys in the same hydrothermal system (Summit & Baross, 2001). In addition, the subseafloor samples showed three phylogenetic groups whereas the sulfide associated samples showed five phylogenetic groups, suggesting that *Thermococcus* that inhabit subseafloor niches have diverged into populations with specific adaptations corresponding to the zone of the subseafloor that they inhabit. The existence of a unique

subseafloor population is further supported by physiological differences observed between the two groups. Samples from the sulfide chimney produced more proteases and could grow under higher zinc ion concentrations, temperature ranges, and salinity gradients than the subseafloor samples. Together, these phylogenetic and physiological differences indicate potential adaptations to an ephemeral sulfide chimney environment, which contrasts with the more static subseafloor environment. The presence of distinct groups within a hydrothermal system where there is mixing of fluids between habitats implies a strong habitat-related selective pressure which could lead to the formation of habitat-specific ecotypes within *Thermococcus*.

Culture dependent techniques can allow for the growth of microbes with specific metabolic capacities but are insufficient when exploring an organism's total metabolic potential. Genomic techniques facilitate this exploration and can help illuminate the potential roles of microorganisms in their environments. The comparison of several genomes of closely related microorganisms to find metabolic adaptations unique to an individual strain or group of strains can be accomplished using a pangenome. A pangenome summarizes the full collection of genes for closely related microorganisms. The core of the pangenome is composed of required metabolic genes found in every genome of the group of microorganisms. For instance, genes for growth, reproduction, and homeostasis are part of the core genome. The remaining genes in the pangenome belong to the variable genome, which is composed of genes that are not found in every member of the group and are not essential for growth but do provide selective advantage, such as genes for alternate metabolic pathways (Tettelin et al., 2005; Medini et al., 2005; Rouli et al., 2015). A pangenome can therefore be used as a tool for examining niche adaptations. For example, genes coding for adaptations that allow a microorganism to survive in a deep-subsurface fluid reservoir for extended period of time are found in the variable genome.

A recent study on *Thermococcus* biogeography using multi locus sequence typing (MLST) and amplified fragment length polymorphism (AFLP) data characterized ninety isolates from vent sites throughout the Pacific Ocean, which resulted in at least ten distinct lineages, including one that contained the five isolates from the 1996 Gorda Ridge megaplume events (Price et al., 2015). The Gorda Ridge lineage also included the type strain *Thermococcus onnurineus* NA1, which was collected from the Manus Basin of the PACMANUS vent field via multiple corer at a depth of 1,650 m (Bae et al., 2006). This strain possesses lithotrophic adaptations, the most noteworthy being the carbon monoxide dehydrogenase (CODH) gene cassette which allows the use of CO as a carbon and energy source (i.e. carboxydotrophy; Lee et al., 2008), therefore providing a growth advantage given the ubiquity of CO in hydrothermal fluids (Symondst et al., 1994). The presence of lithotrophic adaptations along with phylogenetic dissimilarity to other *Thermococcus* thought to originate at the surface of other hydrothermal systems provide evidence that the Gorda Ridge isolates and *T. onnurineus* are part of a deep-subsurface ecotype that was brought to the surface during the 1996 megaplume events.

In order to address *Thermococcus* biogeography on a genomic level, a pangenomic approach was used to compare the genomes of *Thermococcus* isolates spanning ten lineages from seven geographic regions identified by Price et al. (2015), including the deep-subsurface lineage isolated from the aftermath of the 1996 seismic events at Gorda Ridge. It was hypothesized that the variable genomes from the Gorda Ridge isolates would exhibit diverse metabolic pathways that would allow for the use of a variety of carbon and energy sources. Examples of metabolic diversity in *Thermococcus* include the carboxydotrophic adaptations in *T. onnunineus* which allow lithotrophy as an alternative energy source (Lee et al., 2008), and the formate-driven anaerobic respiration that has been observed in the same organism (Kim et al., 2010).

METHODS

Thermococcus isolates and culturing

Isolates were chosen (n=19) to optimally represent each of the ten lineages and seven geographic locations from Price et al. (2015). Isolates were derived from samples collected during research cruises between the years 1988 to 2008 and originated from the Juan de Fuca Ridge, Gorda Ridge, South East Pacific Rise, North East Pacific Rise, Mariana Arc, and Lō‘ihi Seamount (Figure 1). Study sites and methods for serial dilution to extinction for isolation have been previously described (Davis & Moyer, 2008; Huber et al., 2006; Summit & Baross, 2001).

Media formulations, stock solution formulations, and culturing techniques were prepared as previously described (Holden et al., 2001). The medium was transferred into pre-sterilized Balch tubes. The tubes were plugged with pre-sterilized butyl rubber stoppers (Bellco Glass Inc., Vineland, NJ) and sealed with aluminum seals (Bellco Glass Inc.) before the headspace was exchanged with argon gas through a 1 mL syringe fitted with a 0.2 µm filter and 25-gauge needle (Becton Dickinson, Franklin Lanes, NJ) using a gas manifold. A second 25-gauge needle was inserted into the stopper during this process to relieve positive pressure. After gas exchange, filter sterilized 2.5% $Na_2S•9H_2O$ was added as a reducing agent via 25-gauge needle and 1 mL louver lock syringe (Becton Dickinson). A color change from pink to clear was noted before the tubes were inoculated with 0.2 mL of culture using a 25-gauge needle and 1 mL louver lock syringe. The headspace was exchanged with argon gas once again using a 25-gauge needle fitted with a 0.2 µm filter. The pressure relief needle was removed at the end of this process to allow for a slight positive pressure in the tube before incubation. The inoculated Balch tubes were incubated in a sand bath at 70 to 90˚C for 12 to 36 hrs. To confirm growth, tubes were checked with fluorescence microscopy adding 4 µL of 0.25 mM Syto13 nucleic acid stain to four drops of liquid culture.

DNA extraction and analysis

DNA was extracted from the culture tubes using the methods in Price et al. (2015). Freshly grown isolates were transferred to 15 mL centrifuge tubes and centrifuged at 750 × g for 5 min to pellet the sulfur. The supernatant was then transferred to a clean centrifuge tube and centrifuged at 11,000 × g for 10 min in a chilled rotor (4 ˚C). DNA was extracted from this cell pellet using a DNeasy Soil Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. The resulting DNA concentration was determined with a Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA).

Sequencing

DNA from the nineteen *Thermococcus* isolates was sent to the University of Delaware DNA Sequencing & Genotyping Center for library preparation and high throughput sequencing. Libraries were prepared using a Quantabio sparQ Library Prep Kit (Qiagen) following manufacturer's protocol and selecting for a 500 to 1000 base pair (bp) insert size. The libraries were sequenced using paired-end sequencing with 251 cycles per read on an Illumina HiSeq 2500 (Illumina, San Diego, CA).

Genome assembly and quality control

Trimmomatic (v0.38; Bolger et al., 2014) was used to remove the adapter sequences from the raw FASTQ files using the adapter sequences in Supplemental Table 1. To confirm adapter trimming and assess sequence quality, paired output files from Trimmomatic were analyzed with FastQC (v0.11.8; Andrews, 2010). SPAdes (v3.13.0; Bankevich et al., 2012) was used for genome

assembly. Assembly quality was analyzed with QUAST (v5.0.2; Gurevich et al., 2013). Genome data were screened for contaminants and contigs < 200 bp and uploaded to NCBI's GenBank database under the BioProject ID PRJNA523072. Supplemental Table 2 lists accession numbers for each isolate.

Pangenome analysis

Pangenomic comparisons between the Gorda Ridge isolates and the shallow-subsurface isolates were visualized using Anvi'o (v.5.4; Eren et al., 2015) following the workflow described in Delmont & Eren (2018). Genomes were annotated within Anvi'o using NCBI's Clusters of Orthologous Groups (COG) database (Tatusov et al., 2000). When creating the pangenome using the anvi-pan-genome command, NCBI's BLASTp (Altschul et al., 1990) was used for higher accuracy, based on the developer's suggestion. As suggested in the documentation, the inflation parameter for the MCL cluster algorithm (v.14-137; van Dongen, 2000), which is used to identify clusters in amino acid sequence similarity, was changed to ten from the default of two to increase sensitivity. This change was made because the genomes being compared are in the same genus, meaning that they already share high sequence similarity and so the algorithm must be more sensitive to pick up on differences. More distantly related genomes, such as those that are classified on the level of family, would require less sensitivity and therefore a lower MCL inflation parameter. Anvi'o was also used to create an average nucleotide identity (ANI) heatmap based on percent ANI similarity between isolates using the integrated PyANI software (Pritchard et al., 2016).

Functional annotation data were produced using the anvi-get-enriched-functions-per-pan-group program within Anvi'o according to the instructions in the Anvi'o pangenome tutorial

(http://merenlab.org/2016/11/08/pangenomics-v2/#making-sense-of-functions-in-your-pangenome), where groups of samples were defined based on geographic location. The resulting spreadsheet contained a list of COG functions present in the pangenome along with metadata including enrichment scores with associated p-values and corrected p-values, the number of genomes the function occurred in within and outside of the group associated with the COG function, gene cluster ID's associated with the function, and a binary value indicating the presence/absence of the function in the core pangenome.

Enrichment scores represent how unique a COG function is to a particular group. Because functional annotations are done at the gene level, Anvi'o's algorithm first associates each gene cluster with a COG function. In this context, a gene cluster is a group of homologous sequences belonging to one or more genomes as identified by the Anvi'o software based on sequence similarity. If there are multiple functions associated with a single gene cluster, Anvi'o assigns the highest frequency annotation to the COG function. In instances where there is a function that is associated with multiple gene clusters, these gene clusters are all noted for the functional annotation. The latter scenario is more common for distantly related genomes which contain divergent gene clusters with similar functions and is therefore less of an issue for the closely isolates used in this study. Anvi'o builds a frequency table by associating functions with gene clusters as described above, which is used as an input for a functional enrichment test and heuristic analysis to determine which functions are found more frequently in each individual group than would be expected under a normal distribution where each function has an equal probability of occurring in genomes from all groups. This resulting metric, the enrichment score, is calculated using a two sample Z-test to compare its occurrence within a group vs. all other groups. The statistic is rescaled for group size and is more robust for larger groups. However, because it is

applied to many functions within Anvi'o, this metric does not produce a true test statistic and cannot be used for hypothesis testing. Therefore, enrichment scores are intended as a method for sorting data and were implemented in this manner. Positive scores indicate the functional annotation is found more often within the associated group than outside of it, and negative scores are given to functions that are more common outside of the group than within it. To determine which COG annotations are ecologically important functions for the Gorda Ridge group, the COG annotation data for this group were sorted by highest to lowest enrichment score and the highest scoring annotations were chosen for further exploration.

RESULTS

Isolate draft genome statistics

The statistics for the isolate draft genomes are summarized in Table 1. The use of N50 and L50 statistics are an indicator of the contiguity of the assemblies. The N50 statistic is the weighted median point of the contigs. When the contigs are ordered from largest to smallest, N50 is the point that accounts for half of the length of the total genome as expressed in the base pair length for that particular contig. If the contigs ordered from largest to smallest are numbered using integers starting with 1, then L50 is the integer of the N50 contig. Therefore, more contiguous assemblies have larger N50 statistics with smaller L50 statistics because more of the genome sequence is contained within a few large individual contigs.

The 19 isolates had an average genome length of 2,024,709 bp with an average contig size of 32, and an average N50 and L50 of 727,119 bp and 2, respectively. The largest genome was 21S7 from the SE Pacific Rise at 2,368,070 bp in 17 contigs, an N50 of 255,961 bp, and an L50 of 3. The smallest genome was CX2 from the Juan de Fuca Ridge at 1,795,681 bp in 9 contigs, an N50 of 512,482 bp, and an L50 of 2.

Pangenome characteristics

The pangenome for all 19 *Thermococcus* isolates is summarized by a circular plot generated using Anvi'o (Figure 2). The pangenome contains 43,381 genes grouped into 7,530 gene clusters with 10,234 unique COG annotations. The single copy core genome is relatively small in comparison and contains 12,246 genes grouped into 634 gene clusters with 5,565 unique COG annotations. Enrichment scores for the entire pangenome ranged from 3.18 to -3.00. The highest scoring (i.e., enriched) COG was identified as Ligand-binding SRPBCC domain (COG4276),

which was found in the isolate CX2 (Juan de Fuca Ridge). The lowest scoring COGs were part of the core pangenome because, by definition, they are found in every genome in all groups and therefore cannot be enriched in any one group.

Enrichment scores for the Gorda Ridge group ranged from 2.98 to -2.98. The highest scoring function was labeled as a predicted endonuclease distantly related to archaeal Holliday junction resolvase, was found in isolates GR4, GR5, and GR7, and was absent from all other genomes in the pangenome. The highest scoring functional annotations in the Gorda Ridge group had enrichment scores > 2.00. There were twenty-eight high scoring COGs, with nine of these functions found in all four isolates within the Gorda Ridge group (Table 2). These nine COGs result from fifty-two gene calls grouped into ten gene clusters. Their position within the pangenome is indicated on the outer edge of Figure 2.

The dendrogram in Figure 2 organizes the isolates into four lineages which are analogous to the relationships ascertained from the average nucleotide identity (ANI) heatmap (Figure 3). The ANI heatmap relates the isolates in the pangenome to each other based on percentage ANI. The heatmap shows that the Gorda Ridge isolates, along with LS1 from Lō‘ihi Seamount, exhibit a high level of similarity to one another and are moderately similar to isolates from the Mariana Arc, SE Pacific Rise, Mid Atlantic Ridge, and NE Pacific Rise, along with M36 (Lō‘ihi Seamount), and MV11, JDF3, and ES12 from the Juan de Fuca Ridge. The Gorda Ridge isolates and LS1 are dissimilar to MV5 from the Juan de Fuca Ridge, and to M39 and LS2 from Lō‘ihi Seamount. The isolate CX2 from Juan de Fuca Ridge followed the same trend but is slightly more dissimilar to the Gorda Ridge and LS1 isolates than those isolates are to one another. MV5 is the most dissimilar to all isolates in this study. LS2 and M39 share a strong similarity to each other but are dissimilar to all other isolates. 9N3 (NE Pacific Rise) and 21S9 (SE Pacific Rise) are strongly similar to each

other and moderately similar to all other isolates excluding MV5, LS2, and M39. The isolates from the Mariana Arc, Mid Atlantic Ridge, 21S7 and 18S1, 21S9 (SE Pacific Rise), M36, 9N3, and ES12, JDF3, MV11 (Juan de Fuca Ridge) are moderately similar to each other, the Gorda Ridge isolates, LS1, and CX2, but dissimilar to MV5, LS2, and M39.

Anvi'o was also used to examine the genomic relationships exclusive to the four isolates from the megaplume event at Gorda Ridge (Figure 4). The pangenome for this subset of isolates contained 8,815 genes grouped into 2,544 gene clusters. In contrast to the full pangenome in which the single-copy core genome represented a small fraction of the total genes, the single-copy core genome for this subgroup contained the majority of the genes, with 6,060 genes grouped into 1,515 gene clusters with 802 unique COG annotations. The dendrogram shows that GR4, GR5, and GR7 are closely related and unique from GR6, which had a larger section of strain-specific gene clusters than the other Gorda Ridge isolates. In contrast, GR5 had the fewest strain-specific gene clusters and shared the majority of its genes with GR4, which it was most closely related to, and GR7.

Enriched functions in the deep-subsurface lineage

DNA repair. The only functional annotation with an enrichment score > 2.00 that was exclusive to the Gorda Ridge megaplume group was a predicted endonuclease distantly related to Archaeal Holliday junction resolvase (COG0792) which was found in three of the four isolates (GR4, GR5, and GR7). DNA mismatch repair ATPase MutS (COG0249) was found in all four members of this group (GR4, GR5, GR6, and GR7) along with two members of the Lō'ihi Seamount group (M36 and M39). Additionally, three of the four isolates from the megaplume group (GR4, GR5, and GR7) contained several proteins from the clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated proteins (Cas) system: Csm2 small

subunit (COG1421), Csx1 (Csm6) containing CARF domain (COG1517), Csm3 group 7 of RAMP superfamily (COG1337), Csm4 group 5 of RAMP superfamily (COG1567), and Csm5 group 7 of RAMP superfamily (COG1332). Csx1 was also found in LS1 and JDF3, belonging to the Lō‘ihi Seamount and Juan de Fuca groups, respectively. Csm2, Csm3, Csm4, and Csm5 were also found in the LS2 and MVII isolates, belonging to the Lō‘ihi Seamount and Juan de Fuca groups, respectively.

Molecular chaperones. Molecular chaperone DnaK (Hsp70; COG0443) was found in three of the four Gorda Ridge isolates (GR4, GR5, and GR7) and was exclusive to this group. In contrast, chaperonin GroEL (Hsp60 family; COG0459) was part of the core pangenome for all isolates, including those from Gorda Ridge.

Amino acids. The Gorda Ridge megaplume group was enriched for the ABC-type branched-chain amino acid transport system, with COG0411, COG0559 and COG0683 present in GR4, GR5, GR6, and GR7, along with LS2 and M39 from Lō‘ihi Seamount. Additionally, all seven enzymes for the shikimate pathway, including chorismite mutase (COG1605), shikimate 5-dehydrogenase (COG0169), 3-deoxy-7-phosphoheptulonate synthase (DHAP; COG2876), prephenate dehydrogenase (COG0287), 3-dehydroquinate synthase (COG0337), 3-dehydroquinate dehydratase (COG0710), 5-enolpyruvylshikimate-3-phosphate synthase (EPSP; COG0128), and archaeal shikimate kinase (COG1685) were enriched for three of the four isolates from the megaplume group (GR4, GR5, and GR7) as well as LS2 and M39 from Lō‘ihi Seamount.

Molybdenum and tungsten transporters

All isolates possessed a type of molybdenum (Mo) and/or tungsten (W) transporter, and the presence of one over the other appears to be location specific. ModABC enzymes, which are

Mo specific, were absent from the pangenome. However, WtpA (COG0725), which can transport both Mo and W, was present in 13 of the 19 isolates. It was identified in four of the five Juan de Fuca isolates (CX2, MV11, JDF3, and ES12), three of four Lōʻihi Seamount isolates (LS1, LS2, and M39), and all remaining isolates excluding the deep-subsurface group from the Gorda Ridge. WtpB (COG0555) was found in all of the same isolates as WtpA, as well as one Gorda Ridge isolate (GR4). WtpC (COG3839) was absent from the entire pangenome. All four of the Gorda Ridge megaplume isolates were enriched for the ABC-type tungstate transport system (COG2998, COG4662). This annotation corresponds to the TupABC transport system and was also found in one isolate from the Lōʻihi Seamount group (M36) and one isolate from the Juan de Fuca Ridge group (MV5).

DISCUSSION

Hydrothermal vent systems provide a window into the mysteries of the deep-biosphere. The rare opportunity to sequence and compare genomes of *Thermococcus* isolates from the 1996 megaplume eruption at Gorda Ridge has revealed several functional adaptations unique to isolates from the deep-subsurface habitat. These functions largely represent either DNA repair mechanisms, or strategies for amino acid utilization and synthesis.

DNA repair machinery

Double-strand DNA breaks. Holliday junction resolvase Hjc was originally isolated from *P. furiosus* and is an important enzyme in maintaining active double-strand break DNA repair systems in this hyperthermophilic archaeon (Komori et al., 1999), which is also known to be resilient to DNA breakage in environments with high temperatures (DiRuggiero et al., 1997) and ionizing radiation (Peak et al., 1995). A predicted endonuclease distantly related to Archaeal Holliday junction resolvase was enriched for and found exclusively within the megaplume group from Gorda Ridge. The presence of an endonuclease similar to Hjc suggests that the deep-subsurface isolates have prioritized double-strand DNA repair machinery as an adaptation to an extreme environment not encountered by their more shallow subsurface counterparts.

DNA mismatch repair. The DNA mismatch repair (MMR) system finds mismatched base pairs caused by errors in DNA polymerase. This system is highly conserved in Bacteria and Eukarya, and typically involves the MutS enzyme identifying the mismatch in the strand, which is then acted upon by MutL and MutH to excise the incorrect base through endonuclease activity and replace it with the correct base (Acharya et al., 2003). *mutS* was also found within the deep-subsurface group, as well as in two isolates from Lō‘ihi Seamount; however, the pangenome did

not reveal annotations for MutL or MutH enzymes. It has been previously noted that MMR systems are not present in most archaea, although MutS and MutL homologues have been found in *Halobacterium salinarum* NRC-1 and are important for maintaining low mutation rates in this organism, although the exact pathway is unclear (Busch & DiRuggiero, 2010). Therefore, the presence of MutS within the megaplume group may still point to a mechanism of MMR which would function as an important adaptation to hyperthermophilic life in the deep-subsurface.

CRISPR/Cas immune system. The CRISPR/Cas system has recently become of interest for its biotechnology applications; however, it first evolved as a microbial immune system and is found widely throughout the archaea (Haft et al., 2005; Godde et al., 2006, Jinek et al., 2012). The system works by translating short spacer sequences created from past invading genetic elements, such as those from phages, into small CRISPR RNAs (crRNAs) that can locate and eradicate invading nucleic acids with the help of Cas proteins (Garneau, et al. 2010; Hale et al., 2009). The first step in this system is the creation of crRNAs from the spacers, which requires Cas10, Csm2, Csm3, Csm4, Csm5, and Cas6 (Hatoum-Aslan et al., 2011). In type III systems, crRNA intermediates must be further processed to eliminate repeat regions and erroneous spacer sequences to produce a mature crRNA (Delcheva et al., 2011; Hale et al., 2008, Hatoum-Aslan et al., 2011). Hatoum-Aslan et al. (2011) found that Csm2, Csm3, and Csm5 are required for the crRNA maturation process in the type III-A system and operate together to create a ruler mechanism to ensure that the final crRNAs are of the appropriate lengths. A later study elaborated on these findings, discovering that Csm2, Csm3, Csm4, Csm5, and Cas10 form a Cas10•Csm complex, similar to the Cmr complex in *P.furiosus* (Hale et al., 2009), within which mature crRNAs are measured (Hatoum-Aslan et al., 2013).

The CRISPR/Cas proteins Csm2, Csm3, Csm4, Csm5, and Csm6 (Csx1) were enriched in isolates from Gorda Ridge (GR4, GR5, and GR7). Csm2, Csm3, Csm4, and Csm5 were also found in a Juan de Fuca Ridge isolate (MV11) and one Lō‘ihi Seamount isolate (LS2). Csm6 was also found in a Juan de Fuca Ridge isolate (JDF3) and a Lō‘ihi Seamount isolate (LS1). Annotations for Cas6 were found throughout the core pangenome, and Cas10 was found in three Gorda Ridge isolates (GR4, GR5, and GR7), two Juan de Fuca isolates (MV11 and JDF3), LS2 from Lō‘ihi Seamount, MAR from the Mid Atlantic Ridge, Bubble Bath from the Mariana Arc, and 18S1 from the SE Pacific Rise. However, only GR4, GR5, GR7, and LS2 had all seven CRISPR/Cas proteins required for this system. Therefore, only these four isolates have the potential to use the complete type III-A CRISPR-Cas system for protection against invading phage DNA.

The type III-A CRISPR-Cas system has been shown to provide superior protection from infecting phages as compared to type II-A systems. This could indicate that the GR4, GR5, GR7, and LS2 isolates may cope with a higher viral load than the other isolates examined in this study and therefore benefit from a more effective CRISPR immune system that would otherwise decrease their overall fitness. The efficacy of type III-A systems is limited due to increased cell toxicity, as compared to type II-A systems, which contributes to overall reduced fitness when the cell is not actively fighting phage infections (Pyenson et al., 2017, Niewoehner & Jinek, 2017). It has also been observed that type III systems are more abundant in thermophiles because surface modification, another viral defense mechanism, is unavailable due to the more rigid cell walls these organisms must possess to survive higher temperatures (Makarova et al., 2015). Given that deep-subsurface hydrothermal habitats tend to be hotter than the chimney habitats of the same systems, this is a plausible explanation for the enrichment of type III-A CRISPR-Cas enzymes in these isolates.

Molecular chaperones

Molecular chaperones are proteins that ensure the correct folding of other cellular proteins but are not a part of their final structure (Ellis, 1993). The molecular chaperone DnaK (Hsp70) functions as a heat shock protein, stabilizing other proteins within the cell to prevent denaturation and assist in proper folding during times of stress caused by extreme temperatures or other toxic conditions. DnaK is relatively rare in archaea, which typically use the chaperonin Hsp60 (GroEL) instead (Large et al., 2009). This is reflected in this study in that all isolates from all locations contained Hsp60, but DnaK was found only within the megaplume group from Gorda Ridge. DnaK functions in conjunction with the co-chaperones DnaJ (Hsp40) and GrpE (Hartl & Hayer-Hartl, 2002; Zmijewski et al., 2004); however, the two latter proteins were absent from the pangenome. It has been hypothesized that the acquisition of DnaK by select archaeal species occurred via lateral gene transfer from bacteria (Macario & de Macario, 1999). Therefore, the enriched presence of DnaK in the Gorda Ridge lineage may simply be a remnant from horizontal gene transfer and have no remaining function for these isolates. Further research is needed to determine the role, if any, for DnaK in this group of deep-subsurface isolates.

Amino acids

Amino acid catabolism has been observed in many *Thermococcus* species including *Thermococcus litoralis* (Neuner et al., 1990), *Thermococcus kodakaraensis* KOD1 (Fukui et al., 2005), *Thermococcus sibiricus* (Mardanov et al., 2009), and *Thermococcus* strain ES-1 (Ma et al., 1995). This process begins with the oxidative deamination of an amino acid, followed by the oxidative decarboxylation of the resulting 2-oxoacid, and finishing with the hydrolysis of acyl-

CoA coupled to substrate level phosphorylation (Schut et al., 2001). The second step in this pathway can be accomplished with one of several enzymes collectively known as 2-oxoacid:ferredoxin oxidoreductase (KOR) enzymes, reflecting the breadth of metabolic diversity apparent in *Thermococcus* (Blamey & Adams, 1993; Mai & Adams, 1994; Ozawa et al., 2012; Mai & Adams, 1996; Mukund & Adams, 1991). An example of KOR is 2-ketoisovalerate ferredoxin:oxidoreductase (VOR) from *Thermococcus litoralis* which has been proposed to function both as a catalyst in the second step of peptide metabolism and in the biosynthesis of branched-chain amino acids under nutrient poor conditions (Heider et al., 1996).

In some *Thermococcus* species, in the absence of S^0 as a terminal electron acceptor, the final product of the amino acid catabolism pathway is fermented by another ferredoxin oxidoreductase enzyme called aldehyde ferrodoxin:oxidioreductase (AOR), followed by fermentation by alcohol dehydrogenase (Ma et al., 1997; Basen et al., 2014). AOR was first discovered in *P. furiosus* (Mukund & Adams, 1991), and its function was characterized in *Thermococcus* ES-1 (Heider et al., 1995). This enzyme has a tungstopterin structure composed of a [4Fe-4S] cluster surrounding a single tungsten (W) atom. Molybdenum (Mo) and W have similar atomic radii and therefore often play similar roles in molydopterin/tungstopterin enzymes. Although Mo is more common in seawater, W is often found in high concentrations in hydrothermal vent systems, is more stable than Mo at high temperatures (Adams, 1999), and is also less stable in aerobic conditions (Callis & Wentworth, 1977, Kletzin & Adams, 1996, Maia et al., 2016). Therefore, the use of tungstopterin enzymes, including AOR, rather than molybdopterin enzymes are more advantageous in a hot, anoxic deep-subsurface habitat. Substrates specific to AOR include acetaldehyde, isovaleraldehyde, phenylacetaldehyde, and indolealdehyde, which are derived from the oxidation of the amino acids alanine, leucine,

phenylalanine, and tryptophan, respectively (Heider et al., 1995). AOR is part of the core pangenome for the nineteen *Thermococcus* isolates used in this study. This may help explain the abundance of the molybdenum/tungsten transporters from the WtpABC and TupABC systems throughout the pangenome. The WtpABC system which was first discovered in *P. furiosus* and can utilize both W and Mo because it shares a domain with the Mo specific ModABC transporters (Bevers et al; 2006). TupABC, on the other hand, is W specific and was found to be enriched in all four of the deep-subsurface isolates from Gorda Ridge. This reflects an increased need for W over Mo in the deep-subsurface, likely due to increased temperature, decreased oxygen, and a lack of S^0 which would necessitate the use of fermentation following amino acid oxidation given the absence of a terminal electron acceptor.

Given that AOR may be part of an anaerobic pathway that allows the degradation of amino acids for energy, it is therefore appropriate that the Gorda Ridge megaplume group is enriched in W transporters, BCAA transporters, and shikimate pathway enzymes. The shikimate pathway allows for the synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan (Herrmann, 1995). The W transporter TupA brings the nutrient into the cell to be incorporated into the tungstopterin component of AOR. BCAA transporters carry leucine, isoleucine, and valine into the cell, while the shikimate pathway synthesizes aromatic amino acids. The oxidized organic acids of leucine, phenylalanine, and tryptophan may then be oxidized for energy and, in some cases, their products further broken down through fermentation. Similar processes may occur with isoleucine, valine, and tyrosine; however, the oxidized organic acids resulting from these amino acids are not specific to AOR. It could be that these amino acids are less common in the deep-subsurface habitat or, in the case of tyrosine, less easily synthesized in this environment. Given

the extreme limitations of energy production in the deep-subsurface, peptide catabolism and fermentation would be an advantageous adaptation for these microorganisms.

It is interesting to note that the megaplume group was also enriched in molecular chaperone DnaK, which was found to influence the amount of tryptophanase transcribed in *Escherichia coli* (Sieńczyk et al., 2004, Grudniak et al., 2004). Tryptophanase catalyzes the conversion of the aromatic amino acid tryptophan to its indole. Tryptophan is also a substrate specific to AOR and is converted into indolealdehyde by this enzyme. Therefore, the DnaK may not function as a heat shock protein in the Gorda Ridge isolates but rather as a transcription regulator for tryptophanase which could, in turn, affect the synthesis and later degradation, via oxidation and fermentation, of the amino acid tryptophan. If energetically favorable, this process would allow for the storage of energy in the form of aromatic amino acids that are then later catabolized when the cell encounters energy-limited conditions, such as those found in a deep-subsurface fluid reservoir. An examination of tryptophan production, Shikimate enzyme activity, and AOR and alcohol dehydrogenase activity under high and low peptide availability, controlling for S^0 availability, could help clarify this process. If tryptophan is being used for energy storage, high Shikimate enzyme activity and tryptophan production should be observed under high peptide conditions with little AOR and alcohol dehydrogenase activity. When peptide availability is lowered, tryptophan production should decrease along with Shikimate enzyme activity, and higher AOR along with alcohol dehydrogenase activity would be observed as the tryptophan is oxidized and fermented.

Conclusions

This pangenomic analysis has shown that the four *Thermococcus* isolates obtained from the megaplume event at Gorda Ridge in 1996 possess unique adaptations to life in a deep-

subsurface habitat as compared to the remaining fifteen representative isolates originating from shallow subsurface habitats. The most highly enriched deep-subsurface adaptations relate to DNA repair machinery and protein stabilization. The Gorda Ridge group was enriched in a putative endonuclease which likely repairs double-strand DNA breaks, as well as the MutS enzyme which may be part of an MMR system to maintain low mutation rates. Additionally, three of the four members of this group contained elements of a type III-A CRISPR/Cas immune system which is more effective at fighting phage infections than type II systems but can lower the fitness of the organism by increasing cell toxicity. This trade-off may still be beneficial to the organisms if they must contend with a higher viral load than their counterparts from more shallow habitats, or if they are unable to utilize other phage defense mechanisms, such as surface modification, due to living in a high-temperature environment. Taken together, the enriched presence of DNA repair mechanisms suggests that *Thermococcus* in a deep-subsurface habitat may cope with unique stressors such as energy deficiency due to longer residence times in a fluid reservoir, as well as higher heat exposure than the isolates from more shallow habitats. Analogous adaptations are either found sporadically throughout the pangenome, or are completely absent, indicating that these functions are essential to these deep-subsurface organisms.

Additional adaptations of interest relate to this group's potential to use amino acids as an energy source in an otherwise energy-limited environment. The Gorda Ridge isolates are enriched in tungsten transporters, BCAA transporters, and all seven enzymes of the shikimate pathway for aromatic amino acid synthesis. Tungsten is more thermostable, less oxygen tolerant, and is found in higher concentrations in hydrothermal fluid than molybdenum and is also used as the central catalytic site for the AOR enzyme, which is part of the core pangenome. AOR acts on the products of amino acid oxidation in preparation for a final fermentation that occurs in the absence of S^0 as

a terminal electron acceptor. This pathway is therefore useful in environments where S^0 is lacking and peptides are scarce, requiring the microorganism to extract as much energy as possible from what is bioavailable. The Shikimate pathway was also found to be enriched in the Gorda Ridge isolates. This pathway is used to synthesize aromatic amino acids, such as tryptophan. The genomes of these isolates were also enriched in *dnaK*, which has been found to act as a transcription regulator for tryptophanase in *E. coli*. DnaK may act in a similar manner in *Thermococcus*, rather than as a heat shock protein which would require the presence of DnaJ and GrpE, both of which are absent from the pangenome. Assuming it is energetically favorable, the synthesis and later catalysis of tryptophan by oxidation and fermentation may be a method of energy storage used by these isolates when there are no other options available for heterotrophy.

The presence of habitat specific adaptations in the post-megaplume event Gorda Ridge isolates along with their unique phylogenetic placement in relation to isolates from other locations provides evidence that this lineage is diverging from other *Thermococcus* populations from more shallow habitats. These changes are being driven by environmental forcing functions such as higher temperatures and a longer residence time within hydrothermal fluid reservoirs that are lacking in appropriate nutrients and energy inputs for sulfur-dependent heterotrophs. This divergence indicates the formation of a deep-subsurface specific ecotype within *Thermococcus*.

LITERATURE CITED

Acharya, S., Foster, P. L., Brooks, P. & Fishel, R. (2003). The Coordinated Functions of the *E. coli* MutS and MutL Proteins in Mismatch Repair. *Molecular Cell*, 12 (1), 233-246.

Adams, M. W. W. (1999). The biochemical diversity of life near and above 100°C in marine environments. *Journal of Applied Microbiology Symposium Supplement*, 85, 108S-117S.

Alma'abadi, A. D., Gojobori, T., & Mineta, K. (2015). Marine metagenome as a resource for novel enzymes. *Genomics Proteomics Bioinformatics*, 13, 290-295.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology.* 215, 403-410.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Baas Becking, L.G.M. (1934) *Geobiologie of Inleiding Tot De Milieukunde*. W.P. Van Stockum & Zoon, The Hague.

Baker, E. T. (1998). Patterns of event and chronic hydrothermal venting following a magmatic intrusion: new perspectives from the 1996 Gorda Ridge eruption. *Deep Sea Research Part II: Topical Studies in Oceanography*, 45 (12), 2599-2618.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., & Kulikov, A. S. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology.* 19 (5), 455—477.

Bae, Seob, S., Kim, Y. J., Yang, S. H., Lim, J. K., Jeon, J. H. … & Lee, J. (2006). *Thermococcus onnurineus* sp. nov., a Hyperthermophilic Archaeon Isolated from a Deep-Sea Hydrothermal Vent Area at the PACMANUS Field. *Journal of Microbiology and Biotechnology*. 16 (11), 1826-1831.

Basen, M., Schut, G. J., Nguyen, D. M., Lipscomb, G. L., Benn, R. A., Prybol, C. J., … & Adams, M. W. W. (2014). Single gene insertion drives bioalcohol production by thermophilic archaeon. *PNAS*, 1-6.

Bevers, L. E., Hagedoorn, P., Krijger, G. C., & Hagen, W. R. (2006). Tungsten transport protein A (WtpA) in *Pyrococcus furiosus*: the first member of a new class of tungstate and molybdate transporters. *Jornal of Bacteriology*, 188 (18), 6498-6505.

Blamey, J. M. & Adams, M. W. W. (1993). Purification and characterization of pyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*. *Biochimica et Biophysica Acta* 1161, 19-27.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30 (15), 2114–20.

Busch, C. R., & DiRuggiero, J. (2010). MutS and MutL Are Dispensable for Maintenance of the Genomic Mutation Rate in the Halophilic Archaeon Halobacterium salinarum NRC-1. *PLOS*, 5 (2), e9045.

Callis, G. E., & Wentworth, R. A. D. (1977). Tungsten vs. molybdenum in models for biological systems. *Bioinorganic Chemistry*, 7, 57-70.

Chadwick, W. W., Embley, R. W., & Shank, T. M. (1998). The 1996 Gorda Ridge eruption: Geologic mapping, sidescan sonar, and SeaBeam comparison results. *Deep-Sea Research Part II: Topical Studies in Oceanography*, 45 (12), 2547-2569.

Davis, R. E., & Moyer, C. L. (2008). Extreme spatial and temporal variability of hydrothermal microbial mat communities along the Mariana Island Arc and southern Mariana back-arc system. *Journal of Geophysical Research: Solid Earth*, 113 (8), 1-17.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., … Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 471 (7340), 602–607.

Delmont, T. O., & Eren, A. M. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, 6(e4320).

DiRuggiero, J., Santangelo, N., Nackerdien, Z., Jaques, R., & Robb, F. (1997). Repair of extensive ionizing-radiation DNA damage at 95°C in the hyperthermophilic archaeon *Pyrococcus furiosus*. *Journal of Bacteriology*, 197 (14), 4643-4645.

Dick, G. J., Anantharaman, K. Baker, B. J., Li, M., Reed, D. C., & Sheik, C. S. (2013). The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Frontiers in Microbiology*, 4 (121), 1-16.

Dobbs, F. C. & Selph, K. A. (1997). Thermophilic bacterial activity in a deep-sea sediment from the Pacific Ocean. *Aquatic Microbial Ecology*, 13, 209-212.

Ellis, R. J. (1993). The general concept of molecular chaperones. *Philosophical Transactions: Biological Sciences*, 339 (1289), 257-261.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3 (e1391).

Eren, M. A., Shaiber, A., Yousef, M., & Esen, Özcan C. (2016). An anvi'o workflow for microbial pangenomics was published on November 08 - http://merenlab.org/2016/11/08/pangenomics-v2/

Fiala, G., & Stetter, K. O. (1986). *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Archives of Microbiology*, 145 (1), 56–61.

Fox, C. G., & Dziak, R. P. (1998). Hydroacoustic detection of volcanic activity on the Gorda Ridge, February–March 1996. *Deep Sea Research Part II: Topical Studies in Oceanography*, 45 (12), 2513- 2530.

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., & Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. *Genome Research*. 15(3), 352-63.

Garneau, J. E., Dupuis, M., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P. … & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 468, 67—71.

Godde, J. S., & Bickerton, A. (2006). The Repetitive DNA Elements Called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*, 62 (6), 718–729.

Gold, T. The deep, hot biosphere. (1992). *PNAS*, 89, 6045-6049.

Grudniak, A. M., Nowicka-Sans, B., Maciag, M., & Wolska, K. I. (2004). Influence of *Escherichia coli* DnaK and DnaJ molecular chaperones on tryptophanase (TnaA) amount and GreA, GreB stability. *Folia Microbiologica*, 49 (5), 507-512.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 29 (8), 1072-1075.

Haft, D. H., Selengut, J., Mongodin, E. F., & Nelson, K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLOS: Computational Biology*, 1 (6), e60.

Hale, C., Kleppe, K., Terns, R. M., & Terns, M. P. (2008). Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA*, 14 (12), 2572–2579.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., … & Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, 139 (5), 945–956.

Hartl, F. U., & Hayer-Hartl, M. (2002). Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295 (5561), 1852-1858.

Hantoum-Aslan, A., Maniv, I., & Marraffini, L. A. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *PNAS*, 108 (52), 21218—21222.

Hantoum-Aslan, A., Samai, P., Maniv, I., Jiang, W., & Marraffini, L. A. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *The Journal of Biological Chemistry*, 288 (39), 27888-27897.

Heider, J., Kesen, M. A., & Adams, M. W. W. (1995). Purification, characterization, and metabolic function of tungsten-containing aldehyde ferredoxin oxidoreductase from the hyperthermophilic and proteolytic archaeon *Thermococcus* Strain ES-1. *Journal of Bacteriology*, 177 (16), 4757-4764.

Heider, J., Mai, X., & Adams, M. W. W. (1996). Characterization of 2-ketoisovalerate ferredoxin oxidoreductase, a new and reversible coenzyme A-dependent enzyme involved in peptide fermentation by hyperthermophilic archaea. *Journal of Bacteriology*, 178 (3), 780-787.

Herrmann, K. M. (1995). The shikimate pathway: Early steps in the biosynthesis of aromatic compounds. *The Plant Cell*, 7, 907-919.

Holden, J. F., Takai, K., Summit, M., Bolton, S., Zyskowski, J., & Baross, J. A. (2001). Diversity among three novel groups of hyperthermophilic deep-sea *Thermococcus* species from three sites in the northeastern Pacific Ocean. *FEMS Microbiology Ecology*, 36 (1), 51–60.

Huber, J. A., Butterfield, D. A., & Baross, J. A. (2006). Diversity and distribution of subseafloor *Thermococcales* populations in diffuse hydrothermal vents at an active deep-sea volcano in the northeast Pacific Ocean. *Journal of Geophysical Research: Biogeosciences*, 111 (4).

Jannasch, H. W., Wirsen, C. O., Molyneaux, S. J., & Langworthy, T. A. (1992). Comparative physiological studies on hyperthermophilic archaea isolates from deep-sea hot vents with emphasis on *Pyrococcus* strain GB-D. *Applied and Environmental Microbiology,* 58 (11), 3472-3481.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science,* 337 (6096), 816–821.

Kim, Y. J., Lee, H. S., Kim, H. S., Kim, E. S., Bae, S. S., Lim, J. K. … & Kang, S. G. (2010). Formate-driven growth coupled with H_2 production. *Nature,* 467, 362-356.

Kletzin, A., & Adams, M. W. W. (1996). Tungsten in biological systems. *FEMS Microbiology Reviews,* 18 (1996), 5-63.

Komori, K., Sakae, S., Shinagawa, H., Morikawa, K., & Ishino, Y. (1999). A Holliday junction resolvase from *Pyrococcus furiosus*: functional similarity to *Escherichia coli* RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea. *PNAS,* 96, 8873-8878.

Large, A. T., Goldberg, M. D., & Lund, P. A. (2009). Chaperones and protein folding in the archaea. *Molecular Biology of Archaea,* 37 (1), 46-51.

Lee, H. S., Kang, S. G., Bae, S. S., Lim, J. K., Cho, J. K., Kim, Y. J. … Lee, J. H. (2008). The complete genome sequence of *Thermococcus onnurineus* NA1 reveals a mixed heterotrophic and carboxydotrophic metabolism. *Journal of Bacteriology*, 190 (22), 7491-9.

Lovley, D. R., & Chapelle, F. H. (1995). Deep subsurface microbial processes. *Reviews of Geophysics,* 33 (3), 365.

Lundberg, K. S., Shoemaker, D. D., Adams, M. W. W., Short, J. M., Sorge, J. A., & Mathur, E. J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene*, 108, 1-6.

Lupton, J. E. (1996). A far-field hydrothermal plume from Loihi Seamount. *Science,* 272 (5264), 976-979.

Lupton, E. L., Baker, E. T., Garfield, N., Massoth, G. J., Feely, R. A., Cowen, J. P., … & Rago, T. A. (1998). Tracking the evolution of a hydrothermal event plume with a RAFOS neutrally buoyant drifter. *Science,* 280 (5366), 1052-1055.

Lupton, J. E., Baker, E. T., & Massoth, G. J. (1999). Helium, heat, and the generation of hydrothermal event plumes at mid-ocean ridges. *Earth and Planetary Science Letters,* 171 (1999), 343-350.

Ma, K., Loessner, H., Heider, J., Johnson, M. K. & Adams, M. W. W. (1995). Effects of elemental sulfur on the metabolism of the deep-sea hyperthermophilic archaeon *Thermococcus* strain ES-1: characterization of a sulfur-regulated, non-heme iron alcohol dehydrogenase. *Journal of Bacteriology,* 177 (16), 4748-4756.

Ma, K., Hutchins, A., Sung, S., Adams, M. W. W. (1997). Pyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon, *Pyrococcus furiosus*, functions as a CoA-dependent pyruvate decarboxylase. *Proceedings of the National Academy of Sciences,* 94, 9608-9613.

Macario, A. J. L, & de Macario, E. C. (1999). The archaeal molecular chaperone machine: peculiarities and paradoxes. *Genetics*, 152, 1277-1283.

Mai, X., & Adams, M. W. W. (1994). Indolepyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*. *Journal of Biological Chemistry*, 269 (24), 16726-16732.

Mai, X., & Adams, M. W. W. (1996). Characterization of a fourth type of 2-keto adic-oxidizing enzyme from a hyperthermophilic archaeon: 2-ketogluterate ferredoxin oxidoreductase from *Thermococcus litoralis. Journal of Bacteriology*, 178 (20), 5890-5896.

Maia, L. B., Moura, I., & Moura, J. J. G. (2016). CHAPTER 1:Molybdenum and Tungsten-Containing Enzymes: An Overview. In Hille, R., Schulzke, C., & Kirk, M. L. (Ed.), *Molybdenum and Tungsten Enzymes: Biochemistry* (pp. 1-80). London, UK: Royal Society of Chemistry.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J. … & Koonin, E. V. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*, 13 (11), 722-736.

Mayer, F., & Müller (2014). Adaptations of anaerobic archaea to life under extreme energy limitation. *FEMS*, 38 (2014), 449-472.

McMahon, S., & Parnell, J. (2013). Weighing the deep continental biosphere. *FEMS Microbiology Ecology*, 87 (1), 113–120.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics and Development*, 15 (6), 589–594.

Morris, R. M., Rappe´, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., & Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420, 806-810.

Mukund, S., & Adams, M. W. W. (1991). The novel tungsten-iron-sulfur protein of the hyperthermophilic archaebacterium, *Pyrococcus furiosus* is an aldehyde ferredoxin oxidoreductase. *Journal of Biological Chemistry*, 266, 14208-14216.

Neuner, A., Jannasch, H. W., Belkin, S. & Stetter, K. O. (1990). *Thermococcus litoralis* sp. nov: a new species of extremely thermophilic marine archaebacteria. *Archives of Microbiology*, 153, 205-207.

Niewoehner, O., & Jinek, M. (2017). Specialized weaponry: how a type III-A CRISPPR-Cas system excels at combating phages. *Cell Host & Microbe,* 22, 258-259.

Oslowski, D. M., Jung, J., Seo, D., Park, C. & Holden, J. (2011). Production of hydrogen from α-1,4- and β-1,4-linked saccharides by marine hyperthermophilic archaea. *Applied and Environmental Microbiology,* 77 (10), 3169-3173.

Ozawa, Y., Siddiqui, M. A., Takahashi, Y., Urushiyama, A., Ohmori, D., Yamakura, F., … & Imai, T. (2012). Indolepyruvate ferredoxin oxidoreductase: an oxygen-sensitive iron-sulfur enzyme from the hyperthermophilic archaeon *Thermococcus profundus. Journal of Bioscience and Bioengeneering,* 114 (1), 23-27.

Peak, M. J., Robb, F. T., & Peak, J. G. (1995). Extreme resistance to thermally induced DNA backbone breaks in the hyperthermophilic archaeon *Pyrococcus furiosus*. *Journal of Bacteriology,* 177 (21), 6316-6318.

Pledger, R. J., & Baross, J. A. (1991). Preliminary description and nutritional characterization of a chemoorganotrophic archaeobacterium growing at temperatures of up to 110 isolated from a submarine hydrothermal vent environment. *Journal of General Microbiology,* 137 (1991), 203–211.

Price, M. T., Fullerton, H., & Moyer, C. L. (2015). Biogeography and evolution of *Thermococcus* isolates from hydrothermal vent systems of the Pacific. *Frontiers in Microbiology,* 6, 1–12.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., & Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Analytical Methods*, 8 (1), 12–24.

Pyenson, N. C., Gayvery, K., Varble, A., Elemento, O., & Marraffini, L. A. (2017). Broad targeting specificity during bacterial type III CRISPR-Cas immunity constrains viral escape. *Cell Host & Microbe,* 22, 343-353.

Robb, F. T., & Place, A. R. (1995). *Thermophiles: Archaea a Laboratory Manual*. Plainview, NY: Cold Spring Harbor Laboratory Press.

Rouli, L., Merhej, V., Fournier, P. E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7, 72-85.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K.B., & Erlich, H. A. (1988). Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239, 487-491.

Sapara, R., Bagramyan, & Adams, M. W. W. (2003). A simple energy-conserving system: proton reduction coupled to proton translocation. *PNAS* 100 (13), 7545-7550.

Schut, G. J., Menon, A. L., & Adams, M. W. W. (2001). 2-keto acid oxidoreductases from *Pyrococcus furiosus* and *Thermococcus litoralis*. *Methods in Enzymology*, 331, 144-158.

Sieńczyk, J., Skłodowska, A., Grudniak, A., & Wolska, K. I. (2004). Influence of DnaK and DnaJ chaperones on *Escherichia coli* membrane lipid composition. *Polish Journal of Microbiology,* 53 (2), 121-123.

Summit, M. & Baross, J. A. (1998). Thermophilic subseafloor microorganisms from the 1996 North Gorda Ridge eruption. *Deep-Sea Research II,* 45 (1998), 2751-2766.

Summit, M., & Baross, J. A. (2001). A novel microbial habitat in the mid-ocean ridge subseafloor. *PNAS*, 98 (5), 2158-2163.

Symondst, R. B., Rose, W. I., Bluth, G. J. S., & Gerlach, T. M. (1994). Volcanic-Gas Studies: Methods, Results, and Applications. *Reviews in Mineralogy and Geochemistry*, 30, 1-66.

Takai, K., Sugai, A., Itoh, T., & Horikoshi, K. (2000). *Palaeococcus ferrophilus* gen. nov., sp. nov., a barophilic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. *International Journal of Systematic and Evolutionary Microbiology*, 50, 489-500.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*. 28 (1), 33-36.

Teske, A., Edgcomb, V., Rivers, A. R., Thompson, J. R., Gomez, A. de V., Molyneaux, S. J., & Wirsen, C. O. (2009). A molecular and physiological survey of a diverse collection of hydrothermal vent *Thermococcus* and *Pyrococcus* isolates. *Extremophiles*, 13 (6), 905-915.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L. … & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *PNAS*, 102 (39), 13950-13955.

Valentine, D. L. (2007) Adaptations to energy stress dictate the ecology and evolution of the archaea. *Nature*, 5, 316-323.

van Dongen S., Abreu-Goodger C. (2012) Using MCL to Extract Clusters from Networks. In: van Helden J., Toussaint A., Thieffry D. (eds) Bacterial Molecular Networks. *Methods in Molecular Biology (Methods and Protocols)*, vol 804. Springer, New York, NY.

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, 301, 976—798.

Whitaker, R. J. (2006). Allopatric origins of microbial species. *Philosophical Transactions of the Royal Society*, 361, 1975-1984.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *PNAS*, 95, 6578-6583.

Zillig, W., Holz, I., Janekovic, D., Schäfer, W., & Reiter, W. D. (1983). The Archaebacterium *Thermococcus celer* Represents, a Novel Genus within the Thermophilic Branch of the Archaebacteria. *Systematic and Applied Microbiology*, 4 (1), 88-94.

Żmijewski, M. A., Macario, A. J. L., & Lipińska, B. (2004). Functional similarities and differences of an archaeal Hsp70 (DnaK) stress protein compared with its homologue from the bacterium *Escherichia coli. Journal of Molecular Biology*, 336, 539-549.

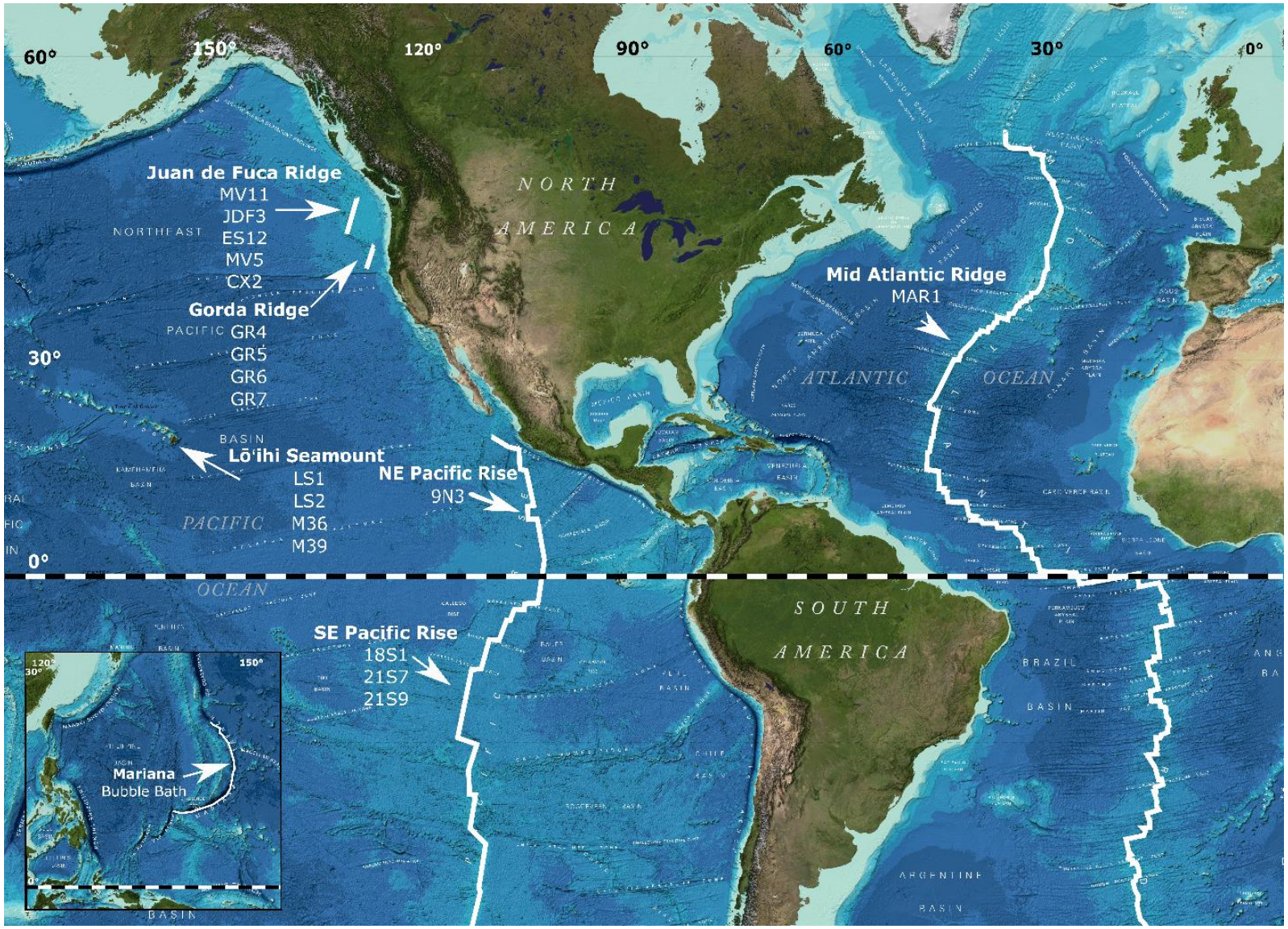
60°
150°
120°
90°
60°
30°
0°
Juan de Fuca Ridge
MV11
JDF3
ES12
MV5
CX2
Gorda Ridge
GR4
GR5
GR6
GR7
NORTHEAST
PACIFIC
30°
0°
NORTH
AMERICA
Mid Atlantic Ridge
MAR1
ATLANTIC
OCEAN
BASIN
Lō'ihi Seamount
LS1
LS2
M36
M39
NE Pacific Rise
9N3
PACIFIC
OCEAN
SE Pacific Rise
18S1
21S7
21S9
SOUTH
AMERICA
BRAZIL
BASIN
ARGENTINE
BASIN
Mariana
Bubble Bath
BASIN

Figure 1. Vent system locations for the nineteen *Thermococcus* isolates used in this study. Site names are identified in bold and the isolates from that location are listed below. Image reproduced from the GEBCO world map, www.gebco.net.

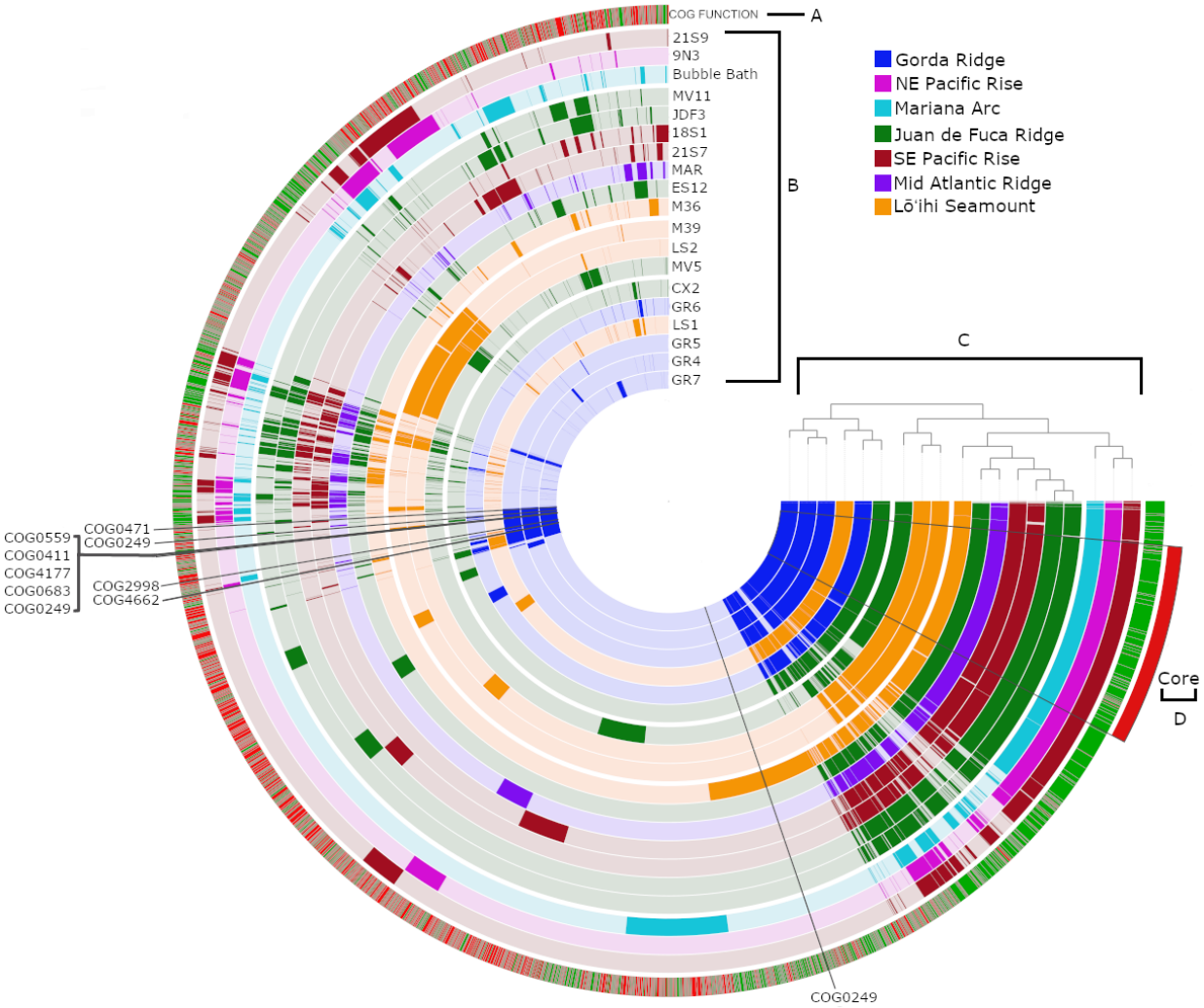
COG FUNCTION
A
21S9
9N3
Bubble Bath
MV11
JDF3
18S1
21S7
MAR
ES12
M36
M39
LS2
MV5
CX2
GR6
LS1
GR5
GR4
GR7
B
Gorda Ridge
NE Pacific Rise
Mariana Arc
Juan de Fuca Ridge
SE Pacific Rise
Mid Atlantic Ridge
Lō'ihi Seamount
C
COG0471
COG0249
COG0559
COG0411
COG4177
COG0683
COG0249
COG2998
COG4662
COG0249
Core
D

Figure 2. Anvi'o pangenome plot for *Thermococcus* isolates (n=19) containing 43,381 genes, 7,530 gene clusters, and 10,234 COGs. A) Availability of COG annotations in the NCBI database for a given gene with green and red indicating the presence or absence of a COG, respectively. COGs with enrichment scores > 2.00 that are common to all four Gorda Ridge isolates are also shown outside the rings. B) Each ring represents the genome of a single isolate. Shaded regions of a ring represent the presence of a gene cluster, defined as a group of homologous sequences belonging to one or more genomes as identified by Anvi'o based on sequence similarity. C) Dendrogram relating isolates based on the similarity of gene cluster frequencies among genomes. D) Single copy core genome containing 12,246 genes, 634 gene clusters, and 5,565 COGs.

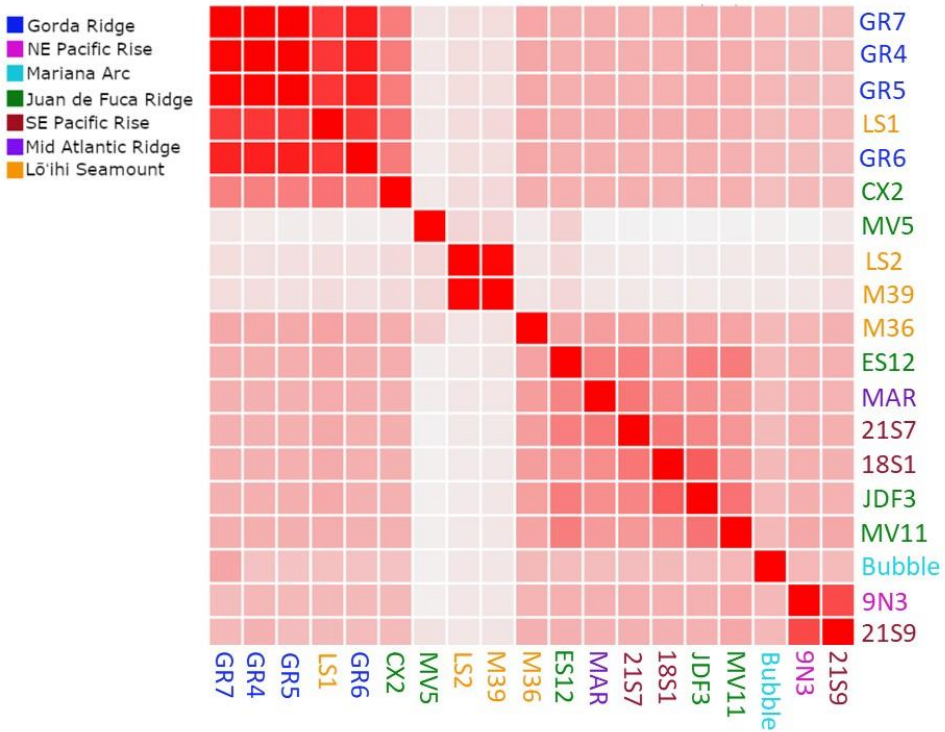
Gorda Ridge
NE Pacific Rise
Mariana Arc
Juan de Fuca Ridge
SE Pacific Rise
Mid Atlantic Ridge
Lō'ihi Seamount
GR7
GR4
GR5
LS1
GR6
CX2
MV5
LS2
M39
M36
ES12
MAR
21S7
18S1
JDF3
MV11
Bubble
9N3
21S9

Figure 3. Percentage average nucleotide identity (ANI) heatmap for *Thermococcus* isolates (n=19) created within Anvi'o using PyANI. Values range from 70% (white) to 100% (red) identity similarity across genomes.

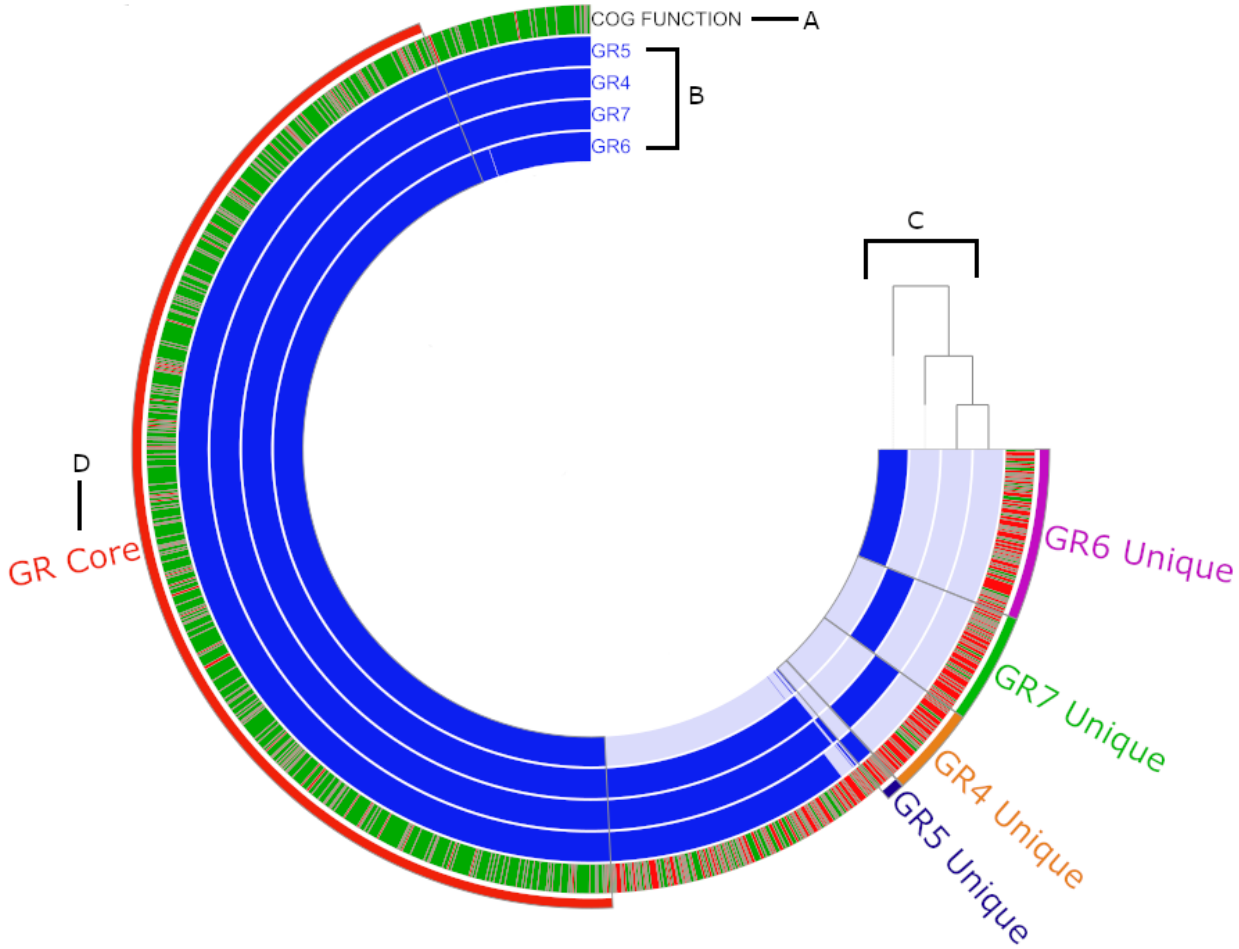
COG FUNCTION
A
GR5
GR4
GR7
GR6
B
C
D
GR Core
GR6 Unique
GR7 Unique
GR4 Unique
GR5 Unique

Figure 4. Anvi'o plot for *Thermococcus* isolates from the Gorda Ridge (n=4) containing 8,815 genes and 2,544 gene clusters. A) Availability of COG annotations in the NCBI database for a given gene with green and red indicating the presence or absence of a COG, respectively. B) Each ring represents the genome of a single isolate. Shaded regions of a ring represent the presence of a gene cluster, defined as a group of homologous sequences belonging to one or more genomes as identified by Anvi'o based on sequence similarity. Strain-specific gene clusters unique to an individual isolate are also identified. C) Dendrogram relating isolates based on the similarity of gene cluster frequencies between genomes. D) Single copy core genome for this subset of isolates containing 6,060 genes, 1,515 gene clusters, and 802 COGs.

Table 1. Summary of genome assembly data from QUAST (v5.0.2) for *Thermococcus* isolates (n=19).

Isolate	# contigs	Total length (bp)	N50	L50	Location	# Gene calls	# COG functions
9N3	10	1,988,207	1,229,293	1	NE Pacific Rise	2,213	3,404
18S1	3	1,996,045	1,989,976	1	SE Pacific Rise	2,104	3,292
21S7	17	2,368,070	255,961	3	SE Pacific Rise	2,585	3,878
21S9	24	1,901,945	1,870,011	1	SE Pacific Rise	2,207	3,382
Bubble Bath	26	2,033,901	357,032	2	Mariana Arc	2,324	3,478
CX2	9	1,795,681	512,482	2	Juan de Fuca Ridge	1,957	3,152
ES12	38	1,962,164	581,465	2	Juan de Fuca Ridge	2,303	3,580
GR4	123	2,020,650	76,437	7	Gorda Ridge	2,364	3,572
GR5	22	1,933,190	198,044	4	Gorda Ridge	2,147	3,342
GR6	3	1,815,275	1,111,285	1	Gorda Ridge	1,975	3,108
GR7	67	1,965,426	326,930	3	Gorda Ridge	2,337	3,528
JDF3	18	2,131,792	377,999	3	Juan de Fuca Ridge	2,345	3,560
LS1	7	2,000,362	618,524	2	Lō‘ihi Seamount	2,211	3,436
LS2	28	2,292,563	323,104	3	Lō‘ihi Seamount	2,519	3,968
M36	129	2,061,277	1,390,874	1	Lō‘ihi Seamount	2,654	4,004
M39	26	2,271,597	466,036	3	Lō‘ihi Seamount	2,506	3,966
MAR	8	2,009,736	1,390,887	1	Mid Atlantic Ridge	2,277	2,277
MV5	45	1,988,329	304,109	3	Juan de Fuca Ridge	2,259	3,584
MV11	7	1,933,257	434,813	2	Juan de Fuca Ridge	2,094	3,204
					Totals:	43,381	65,715

Table 2. Enriched COG functions (enrichment score > 2.00) for *Thermococcus* isolates from Gorda Ridge. Enrichment scores were calculated within Anvi'o representing how unique a COG function is to a group by comparing its occurrence within a group vs. all other groups. Positive scores indicate the function is found more often within the group than outside of it. Functions that are part of the core pangenome are excluded from this table. Functions in bold were common to all four isolates within the Gorda Ridge group.

COG Function	COG Accession	Enrichment Score	Gene Cluster IDs
Predicted endonuclease distantly related to archaeal Holliday junction resolvase	COG0792	2.98	GC_00002748
Di- and tricarboxylate transporter	**COG0471**	**2.98**	**GC_00002194**
Molecular chaperone DnaK (HSP70)	COG0443	2.98	GC_00002788
Shikimate 5-dehydrogenase	COG0169	2.49	GC_00002190
ABC-type branched-chain amino acid transport system, ATPase component	**COG0411**	**2.49**	**GC_00001993, GC_00001979**
CRISPR/Cas system CSM-associated protein Csm2, small subunit	COG1421	2.49	GC_00005862, GC_00002776, GC_00006343
Chorismate mutase	COG1605	2.49	GC_00002201
ABC-type tungstate transport system, permease component	**COG2998**	**2.49**	**GC_00001976**
Transketolase, N-terminal subunit	COG3959	2.49	GC_00001891
3-deoxy-D-arabino-heptulosonate 7-phosphate (DAHP) synthase	COG2876	2.49	GC_00002223
CRISPR/Cas system-associated protein Csx1, contains CARF domain	COG1517	2.49	GC_00005033, GC_00002769, GC_00004997
ABC-type branched-chain amino acid transport system, periplasmic component	**COG0683**	**2.49**	**GC_00001790**
DNA mismatch repair ATPase MutS	**COG0249**	**2.49**	**GC_00002169, GC_00002830, GC_00006162**
Prephenate dehydrogenase	COG0287	2.49	GC_00002785, GC_00003432
3-dehydroquinate synthetase	COG0337	2.49	GC_00002189
Chorismate synthase	COG0082	2.49	GC_00002147
3-dehydroquinate dehydratase	COG0710	2.49	GC_00003380, GC_00002802
5-enolpyruvylshikimate-3-phosphate synthase	COG0128	2.49	GC_00002756, GC_00003368
Branched-chain amino acid ABC-type transport system, permease component	**COG0559**	**2.49**	**GC_00001971**
CRISPR/Cas system CSM-associated protein Csm3, group 7 of RAMP superfamily	COG1337	2.49	GC_00004136, GC_00002778
ABC-type branched-chain amino acid transport system, permease component	**COG4177**	**2.49**	**GC_00002021**
CRISPR/Cas system CSM-associated protein Csm5, group 7 of RAMP superfamily	COG1332	2.49	GC_00002750, GC_00005139, GC_00006021
ABC-type tungstate transport system, periplasmic component	**COG4662**	**2.49**	**GC_00002016**
CRISPR/Cas system CSM-associated protein Csm4, group 5 of RAMP superfamily	COG1567	2.49	GC_00007234, GC_00006196, GC_00002744
Archaeal shikimate kinase	COG1685	2.49	GC_00002206
Predicted DNA-binding transcriptional regulator YafY, contains an HTH and WYL domains	COG2378	2.11	GC_00002898
Uncharacterized conserved protein, contains ParB-like and HNH nuclease domains	COG1479	2.10	GC_00006479, GC_00002730, GC_00004271, GC_00007041
Transketolase, C-terminal subunit	COG3958	2.10	GC_00006940, GC_00003551, GC_00001901, GC_00003420

Supplemental Table 1. Adapter sequences trimmed using Trimmomatic (v0.38).

Description	Sequence
TruSeq Adapter Index 1	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 2	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 3	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 4	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 5	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 6	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 7	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA
TruSeq Adapter Index 8	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 9	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 10	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 11	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 12	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAA
TruSeq Adapter Index 13	GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 14	GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 15	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 16	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 18	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTCCGCACATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 20	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTGGCCTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 21	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTTTCGGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 22	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTACGTAATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 25	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
TruSeq Adapter Index 27	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTCCTTTATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA
Illumina Single End PCR Primer 1	GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA

Supplemental Table 2. *Thermococcus* isolate (n=19) GenBank accession numbers for BioProject ID PRJNA523072.

Isolate name	**GenBank accession number**
9N3	SNUV00000000
18S1	SNUU00000000
21S7	SNUT00000000
21S9	SNUS00000000
Bubble Bath	SNUR00000000
CX2	SNUQ00000000
ES12	SNUP00000000
GR4	SNUO00000000
GR5	SNUN00000000
GR6	SNUM00000000
GR7	SNUL00000000
JDF3	SNUK00000000
LS1	SNUJ00000000
LS2	SNUI00000000
M36	SNUH00000000
M39	SNUG00000000
MAR	SNUF00000000
MV5	SNUE00000000
MV11	SNUD00000000