UNIVERSITY OF CALIFORNIA, SAN DIEGO

Uncovering Genomic Properties of Microbial Life in the Deepest Portions of the Atlantic and Pacific Oceans One Cell at a Time

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Marine Biology

by

Rosa Iris Leon-Zayas

Committee in charge:

Professor Douglas H. Bartlett, Chair
Professor Eric E. Allen
Professor Lihini I. Aluwihare
Professor Farooq Azam
Professor Terence T. Hwa
Professor Roger S. Lasken

2014

The Dissertation of Rosa Iris Leon-Zayas is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

DEDICATION

To my family for been a source of unconditional support:

Mami, Papi, Pablito, Mari y Mamá,

But most of all to Jennifer and Scarla

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BLAST	basic local alignment search tool
COG	cluster of orthologous groups
CP	candidate phyla
DOC	dissolved organic carbon
DNA	deoxyribonucleic acid
EMP	Embden–Meyerhof–Parnas
FACS	Fluorescence-activated cell sorting
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDA	Multiple Displacement Amplification
MPa	Megapascal (0.101 MPa = 1 atmosphere)
MCRG	most closely related genome
NAD(P)	Nicotinamide adenine dinucleotide (phosphate)
ORF	open reading frame
OTU	operational taxonomic unit
PCR	polymerase chain reaction
PP	pentose phosphate pathways
PFK	phosphofructokinase
POM	particulate organic matter
RNA	ribonucleic acid
rRNA	ribosomal RNA
SAG	Single amplified genome
TCA	tricarboxylic acid cycle
tRNA	transfer RNA

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is with my whole heart that I thank the countless number of people that made the completion of this dissertation possible. First and for most, I would like to thank my advisor Doug Bartlett. Thanks to his wisdom (scientific and other wise), he has guided me to pursue new challenges and to conquer amazing opportunities that I had never though I could have accomplished. His scientific critical thinking, is with out doubt, the role model I will strive to emulate for the rest of my scientific career. To the rest of my committee Lihini, Eric, Roger, Farooq and Terry, thank you for your support, for opening your laboratories (and lab meetings) to me and provide your ideas, advise and scientific expertise. It has been an honor to have the opportunity of learning from such an extraordinary group of scientists.

I would like to thank the past and present members of the Bartlett Lab, especially Logan Peoples for being my partner in crime throughout the last two years. Thanks to the Allen, Farooq, Aluwihare and Lasken's labs for opening their doors to me. I would also like to especially thank Roger Chastain for his invaluable insight of high pressure equipment, to the National Geographic engineering team Erik Berkenpas, Mike Shepard and Graham Wilhelm for constructing and facilitating the use of their drop-camera free falling vehicle, which I used to collect much of the samples used in this dissertation. I would also like to sincerely thank the UC-Ship funds and all the technical personnel from Ship Operation and Marine Technical Support, which gave me the opportunity and helped me lead my first oceanographic expedition. Along those lines I most thank Jenan Kharbush for her assistance in planning the expedition. Lastly, all of this work would

have not been possible without the incredible amounts of help from Dr. Roger Lasken and Mark Novotny, who helped me learn the traits of single cell genomics and encouraged my many hours of single cell sorting, amplification and preparation for whole genome amplification.

People say it takes a village to complete a PhD and I've had the support of two "big extended families" through the past seven years. My SIO family of graduate students and post-docs, especially the 2007 grad-student cohort, who were there for me to share their science and their lives. Also my non-SIO family, the San Diego Womens Chorus, which provided a musical outlet and reminded me to appreciate the beautiful things in life. To them, I also will always owe to have met the most important person in my life my invaluable friend, partner and soul mate Jennifer Crick. She has read every page of my dissertation and now understands words like piezophile, for her and all of what she has sacrificed by my side I will be eternally grateful. Finally, to my family in Puerto Rico who raised me to never quit and always smile, thank you.

Chapter 2 is a full-length manuscript submitted for publication: Rosa León Zayas, Mark Novotny, Sheila Podell, Charles M. Shepard, Eric Berkenpas, Sergey Nikolenko, Pavel Pevzner, Roger S. Lasken and Douglas H. Bartlett. 'Microbial Metabolic Properties at Greater Than 8,000 Meters Depth Within the Puerto Rico Trench Inferred From Single Cell Genomics' with permission from all coauthors

Chapter 3 is a full-length manuscript in preparation for publication: Rosa León Zayas, Logan Peoples, Sheila Podell, Mark Novotny, James Cameron, Roger S. Lasken and Douglas H. Bartlett. 'Expansion of the metabolic potential of candidate phylum OD1

based on cells obtained from the Challenger Deep, Mariana Trench' with permission from all coauthors

Chapter 4 is a full-length manuscript in preparation for publication: Rosa León Zayas, Logan Peoples, Jonathan Tarn, Sheila Podell, Mark Novotny, James Cameron, Roger S. Lasken and Douglas H. Bartlett. 'Genomic characterization of Marinimicrobia (Marine Group A, SAR406) single cell genomes from the Challenger Deep' with permission from all coauthors

VITA

2007	B.S. University of Puerto Rico, Mayaguez Campus Industrial Biotechnology Program (Magna cum laude)
2010	M.S. University of California, San Diego, Scripps Institution of Oceanography, Oceanography
2014	Ph.D. University of California, San Diego, Scripps Institution of Oceanography, Marine Biology
2011 – 2013	Marine Biotechnology Training Program, NIH fellow, Scripps Institution of Oceanography, La Jolla CA
2009 – 2014	NSF Pre-Doctoral Graduate Research Fellow, Scripps Institution of Oceanography, La Jolla CA
2008 – 2009	NSF Graduate Teaching Fellows in K-12 Education (GK-12), Scripps Institution of Oceanography, La Jolla CA
2007 – 2008	Alliance for Graduate Education and the Professoriate (AGEP) First Year Fellow, Scripps Institution of Oceanography, La Jolla CA, 2007 – 2008

ABSTRACT OF THE DISSERTATION

Uncovering Genomic Properties of Microbial Life in the Deepest Portions of the Atlantic and Pacific Oceans One Cell at a Time

by

Rosa Iris Leon-Zayas

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2014

Professor Douglas H. Bartlett, Chair

This dissertation presents the analyses of twenty-eight single amplified genomes (SAGs) distributed among four major phyla or candidate phyla of archaea and bacteria: *Thaumarchaeota, Proteobacteria, Parcubacteria* and *Marinimicrobia*. Samples were obtained from 8,219 m and 10,908 m depth within the hadal ecosystems of the Puerto Rico Trench (PRT) and Challenger Deep (CHDE) portion of the Mariana Trench, respectively, and microbes associated with seawater,

invertebrates and surficial sediments were sorted, amplified by multiple displacement amplification and sequenced using the HiSeq 2000 Illumina platform. Assembled and annotated genomes were analyzed and compared to genomes derived from closely related microbes from other habitats with the goal of understanding PRT and CHDE microbes' metabolic adaptations to deep-sea conditions.

Four single amplified genomes (SAGs) were recovered from the PRT: PRT *Nitrosopumilus*, PRT SAR11, PRT *Marinosulfonomonas,* and PRT *Psychromonas*. These microbes are all members of deep-sea phylogenetic clades. The PRT *Nitrosopumilus*, possesses genes associated with mixotrophy, including those associated with lipoylation and the glycine cleavage pathway, and remarkably, may possess the ability to produce fatty acids and lipids. PRT SAR11 encodes for glycolytic enzymes previously reported to be missing in this highly abundant and cosmopolitan group. The PRT *Marinosulfonomonas and* PRT *Psychromonas* SAGs possess genes that may supplement their energy demands through nitrous oxide and hydrogen oxidation.

From the CHDE, 13 SAGs were analyzed that belong to the candidate phylum *Parcubacteria* (OD1), a group of uncultivated microbes characterized by reduced genomes with limited metabolic potential. Comparative genomics was used to examine the metabolic potential harbored by *Parcubacteria* SAGs (OD1-DSC). Horizontally transferred genes were abundant in the *Parcubacteria* genomes especially for genes laterally transferred from members of the archaea. Results indicated that some OD1 cells are capable of much greater metabolic versatility and genetic exchange than previously ascribed to this candidate phylum.

The other candidate phylum analyzed as part of this dissertation is the Marinimicrobia (Marine Group A, SAR406), which has been suggested to be an abundant contributor in deep ocean microbial communities, but information about their metabolism and physiology remains minimal. Six Marinimicrobia SAGs were recovered and their phylogenetic and metabolic characteristics were explored. Results revealed two distinct Marinimicrobia clades not associated with previously described Marinimicrobia phylogenetic groups, but mostly associated with sequences obtain from deep-sea environments, particularly sediments. Bioinformatic analyses indicated that Marinimicrobia SAGs take advantage of carbon monoxide and reduced sulfur compounds to supplement their energy requirements. Osmotic and oxidative stress regulation were also found to be over abundant in the hadal Marinimicrobia. Taking all of these genome studies into consideration it is hypothesized that diversified carbon and energy acquisition pathways are a hallmark of many hadal microbes, along with enhanced osmotic pressure adaptation, perhaps as a means to counteract extreme hydrostatic pressure. In all cases this research has expanded the currently available knowledge of the phylogenetic placement and metabolic potential of the microbial groups studied.

CHAPTER 1

From inception to contemporary insight: The evolution and application of current understandings and practices with regard to deep-sea microbes

The evolution of deep-sea microbiology:

Since the era of Edward Forbes and his azoic hypothesis of a life depleted ocean below 600 m, much has changed with regard to the study of microbial life in the deep ocean (Jannasch and Taylor, 1984). The Challenger expedition, in the late 19th century (1873-1876), is considered the beginning of deep-sea biology as a field and it paved the way for scientist like Certes to investigate microbial life at great ocean depths (Certes, 1884). In the mid 20th century the field of deep-sea microbiology was invigorated with the contributions of scientists like Claude ZoBell, A. Aristides Yayanos, Holger Jannasch and Rita Colwell (Prieur and Marteinsson, 1998). Until that point there were only indications that bacteria could inhabit the deep ocean and that they could live at high pressure, as demonstrated by cell counts using simple microscopes and colonies on agar plates (Fischer, 1894). ZoBell and Johnson (1949) contributed significantly to the growth of deep-sea microbiology with their exploration of the effects of pressure and temperature on bacterial cultures. They concluded that microorganisms isolated from deep-sea environments grew more readily under pressure than terrestrial isolates. The name "barophile" was introduced to describe those microorganisms, hypothesized but not yet proven to exist, that grow and metabolize more efficiently at pressures greater than atmospheric pressure.

In 1979, A. A. Yayanos isolated the first barophilic bacterial species (Yayanos *et al*, 1979). The isolation of *Psychromonas sp*. CNPT3 substantiated the views advanced by ZoBell that deep-sea microorganisms had specific adaptations for improved growth at high pressure and low temperature. Yayanos then isolated the first obligate barophile (referring to a barophile that is unable to grow at atmospheric pressure), *Colwellia sp*.

MT41, from an amphipod collected in the Mariana Trench (Yayanos *et al*, 1981). From this study Yayanos and his group concluded that pressure is an important parameter for zonation along the water column, and that reproduction rates at depth are slow. In 1995 Yayanos proposed the name "piezophile" to replace the name barophile, describing those organisms that grow better at high pressure than at atmospheric pressure. It seemed more appropriate due to the meaning of the prefixes: "baro" means "weight", while "piezo" means "pressure" (Yayanos, 1995). The isolation of obligate piezophiles, piezophiles and piezotolerant organisms lead to their taxonomic identification and analyses of their specific mechanisms of high pressure adaptation. It is generally accepted that high pressure-adapted organisms are very similar to their surface water relatives, suggesting that the evolutionary changes for pressure adaptation do not involve dramatic genetic alterations (Bartlett, 2002).

Deep ocean environmental conditions:

Deep ocean environmental conditions are extreme when compared to surface conditions. The deep sea is generally characterized by the lack of sun light past the first few hundred meters of the water column, near freezing temperatures and increased proportions of recalcitrant organics. Researchers hypothesize that these characteristics give rise to unique microbial communities that have adapted to such conditions (Oger and Jebbar, 2010). The term hadal zone is used to described ocean regions with a depth greater than 6,000 meters. They are found primarily in the deepest trenches of the ocean. Among the factors that distinguish hadal trenches from other deep-sea environments is their biological diversity. They contain specialized fauna distinct from shallower deep-

sea fauna and within trenches vertical zonation exists (Blankenship *et al*, 2006; France, 1993). Species of hadal fauna are restricted to single or adjacent trenches and as a result trenches have been referred to as "zoographic provinces" (Vinogradova, 1997). At this time it is not clear how different trench microbial communities are from one another or from those present in bathypelagic and abyssal regions.

In deep-sea environments the absence of sunlight forces microorganisms to acquire carbon and energy from either exported production (White, 2009), or through chemoautotrophy. Heterotrophic prokaryotic organisms utilize exclusively dissolved organic matter (DOM) for metabolic processes, either by directly taking up the dissolved forms or by using enzymes to break down particulate organic matter (POM) to DOM (Azam & Malfati, 2007). The ocean's DOM and POM pool is more biochemically available at surface than at depth in the ocean (Aluwihare *et al*, 2002). Throughout the water column and in the sediments organisms have adapted to metabolize organic matter available to them and recycle essential nutrients that fuel the biogeochemical cycles (Aristegui *et al*, 2009).

It has been reported that the carbon fixed in the deep ocean is comparable to as little as 15% and as much as 50% of the amount of carbon exported from the photic zone (Reinthaler *et al* 2010). Aerobic ammonia oxidation has become one the most thoroughly studied processes for carbon fixation in recent years because of the ubiquity of ammonia oxidizing Thaumarchaeaota in deep waters, accounting for 60% (in the Atlantic) to 81% (in the Pacific) of the total carbon fixation (Walker *et al,* 2010; Swan *et al*, 2011). Members of the division Thaumarchaeaota are among the most abundant archaea on the planet (Pester *et al*, 2011). Characterized by their ability to oxidize ammonia

autotrophically, members of the division Thaumarchaeota have been suggested to play a major role in the nitrogen and carbon cycle, particularly in the deep ocean (Konstantinidis *et al*, 2009; Herndl *et al*, 2005). Besides carbon fixation via ammonia oxidation, at depths of ~ 800 m, members from the uncultured groups ARCTIC96BD-19, Agg47 and SAR324 have been reported to possess ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) enzymes indicative of carbon fixation via the Calvin-Benson-Bassham cycle and it was demonstrated that the SAR324 clade can fix carbon coupled to the oxidation of reduced sulfur compounds (Swan *et al*, 2011). Chemoautotrophy is more prevalent in sediments where below a few cm oxygen becomes limiting, creating an anaerobic environment. The small amounts of organic matter leads microorganisms to synthesize their own organic carbon. Once the oxygen is used up, the use of different electron acceptors, such as nitrate or sulphate, becomes more prevalent. The consumption of the small amounts of organic matter under anaerobic conditions, via fermentative or respiratory metabolism, leads to the production of reduced compounds such as H_2, H_2S and ammonia, which can be subsequently used as energy sources for chemoautotrophs. Ocean sediments harbor not only high microbial biomass but great diversity due in part to the varied environmental conditions found throughout, e.g. surficial accumulation of organic mater and oxygen consumption, and development of stratified redox gradients (Torsvik *et al*, 2002; Zinger *et al*, 2011; Edwards *et al,* 2012).

The extent of chemoautotrophy in hadal environments is not well understood and potentially limited, as reflected in the Puerto Rico Trench metagenome (Eloe *et al*, 2011). The ammonia oxidizing Thaumarchaeota described above have also been reported to be prevalent in ultra-deep environments (Eloe *et al*, 2011). Chemoautrophic bacteria and

archaea have also been identified in the Japan Trench cold seep communities (Arakawa *et al,* 2006; Inagaki *et al*, 2002).

Adaptations to deep-sea environments:

High hydrostatic pressure is the most unique environmental characteristic of deep-sea environments. Pressure is a three-dimensional force that increases at a rate of 1 atmosphere per 10 meters of depth. Piezophilic strains like *Photobacterium profundum* strains SS9 and *Shewanella sp*. DSS12 (both subphylum *Gammaproteobacteria)*, as well as mesophilic strains like *Escherichia coli (sub*phylum *Gammaproteobacteria)* and *Saccharomyces cerevisiae (Fungi)* have been used to investigate the genetic traits that confer pressure adaptation. Using these species, processes involved in membrane protein regulation (RseB), membrane transport/nutrient uptake (OmpH), DNA recombination (RecD), cytochrome assembly (CydD) and cell division (FtsZ) have been associated with adaptations to high pressure (Bartlett, 2002). Analyses conducted with SS9 conditional mutants indicate that chromosome partitioning and ribosome function are processes affected by both low temperature and high pressure (Lauro *et al*, 2007). High pressure also appears to influence energy yielding processes, as shown by differences in the respiratory chains of piezophilic isolates grown at atmospheric pressure and high pressure (Simonato *et al*, 2006). Low temperature and high pressure both cause a decrease in cell membrane fluidity (Royer, 1995). Organisms at depth compensate by producing higher quantities of unsaturated fatty acids. Some piezophiles produce long chain omega-3 polyunsaturated fatty acids (eicosapentaenoic acid and docosahexaenoic acid) (Allen and Bartlett, 2002). Monounsaturated fatty acids can also be required for growth at high

pressure (Allen *et al*, 1999). Little is known about protein adaptations in piezophilic microbes; however, studies performed with single-stranded DNA-binding proteins (SSB) do show decreased abundance of proline and glycine in the highly variable region of the proteins (Chilukuri and Bartlett, 1997). This may reflect decreased protein flexibility and increased stability at high pressure. More recently, studies performed using piezophilic bacterium *Shewanella violacea* strain DSS12 have provided much information in terms of protein function and regulation of deep-sea adapted microbes (Kato, 2011). For *S. violacea* respiratory proteins, cell division protein FtsZ, RNA polymerase, dihydrofolate reductase (DHFR), and isopropylmalate dehydrogenase (IPMDH), have all been examined, and some of them are much more stable and active under higher-pressure conditions (Kato, 2011).

Genome sequencing has also been applied to individual genomes derived from cultured deep-sea microbes. The sequencing of whole genomes of piezophiles and their shallow-water relatives has made it possible to compare protein families and metabolic pathways, as in the case of *Photobacterium profundum* strains SS9 (Vezzi *et al*, 2005). Among the genomic characteristic that are suggested to promote SS9 adaptation to life in the deep sea are its large number of ribosomal operons, which reflects SS9's ability to respond to changes in environmental conditions. Adaptation to life at depth and variable nutrient inputs is also evident in the overrepresentation of genes involved in carbohydrate transport and metabolism. Indications of modification in the electron transport chain were also noted in the SS9 genome, as well as genes for assimilatory and dissimilatory reduction of nitrate, tetrathionate, dimethylsulfoxide, fumarate, sulfite, and trimethylamine-N-oxide (TMAO).

Piezophiles have also been targeted for functional genomics. Vezzi and colleagues investigated the transcriptional profile of SS9 using micro array technology. At high pressure a number of unusual metabolic properties were upregulated. Examples include the potential for amino acid fermentation and the respiratory reduction of TMAO. It was also noted that at high pressure metabolic pathways for the degradation of different polymers such as chitin, pullulan, and cellulose were activated. Genes induced at atmospheric pressure were very different from those turned on at elevated pressure and included systems facilitating protein folding, apparently reflecting the stress perceived by a piezophile when shifted to suboptimal atospheric pressure conditions (Vezzi *et al*, 2005). More recently, the SS9 transcriptome was analyzed as a function of pressure using RNA-seq methods (Campanaro *et al*, 2012). Genes under the control of a pressure-responsive transmembrane transcription factor were identified. One of the most significant findings was the identification of large amounts of untranslated regions (UTRs), which pointed out that SS9 harbors high potential for novel cis-regulatory RNA structures, and thus cis-regulatory control of gene expression.

The proteome of SS9 has also been studied, and proteins differentially expressed as a function of pressure noted (Le Bihan *et al*, 2013). Many of these proteins have been previously identified as playing important roles in cellular adaptation (Vezzi *et al*, 2005). However, some of the differentially expressed proteins either have not previously been identified in high-pressure adaptation mechanisms or were not regulated as expected. Proteins up-regulated at high pressure are involved in respiration, ABC-transporters for ions, sugars and amino acids, regulatory and ribosomal proteins, as well as some enzymes involve in the glycolysis pathway and alcohol metabolism (Le Bihan *et al*, 2013).

The understanding of molecular adaptations to high pressure has benefited primarily from genetic manipulation of selected piezophilic species. As a result this understanding is skewed towards a narrow phylogenetic group of microbes that have been isolated from the deep sea, namely members of the *Gammaproteobacteria*. To overcome the "great plate anomaly" (Staley and Konopka, 1985), microbiologists have resorted to culture independent studies to understand those organisms that cannot be cultured (~99.9% of all microbes in the ocean). Studying microbes and whole microbial communities without the need of cultivations has developed within the field of genomics. The term metagenomics was first used by Handelsman and colleagues in 1998 in a study of soil microbes while cloning random environmental DNA (Handelsman *et al*, 1998). The study of DNA from whole microbial communities became more popular at the turn if the 21st century with the analysis of fosmid vectors containing environmental DNA. As an example, one of the most significant first discoveries came from DNA extracted from the Antarctic and deep Pacific ocean with the goal of better characterizing the phylum now classified as Thaumarchaeota (Beja *et al*, 2002). Since then there have been about 45 major metagenomic studies of ocean environments including deep-sea habitats (Gilbert and Dupont, 2011). Metagenomic sequences of deep-sea samples and phylogenetic studies of 16S rRNA genes have been performed in a number of deep-ocean environments. The metagenomic analyses include the 4000 m Hawaii Ocean Time series (HOT) (DeLong *et al*, 2006; Konstantinidis *et al*, 2009), the 3000 m station Km3 (Martin-Cuadrado *et al*, 2007) and 4900 m Matapan-Vavilov Deep (Smedile *et al*, 2013) both in the Mediterranean Sea, and the only hadal metagenome, a seawater sample obtained from 6000 m in the Puerto Rico Trench (Eloe *et al*, 2011). These studies have

revealed that the microbial diversity, gene inventory and metabolic potential are broad in the deep ocean and differ from those present in shallower environments.

From the open-ocean oligotrophic 4000 m Hawaii Ocean Time series (HOT) metagenome certain groups of microbes were discovered to be more abundant at depth than at the surface (e.g. Chloroflexi and Planctomyetes – like sequence), and some metabolic processes and specific genes were likewise differentially distributed (DeLong *et al*, 2006). Among the metabolic properties more represented in the 4000 m sample were glyoxylate and dicarboxylate metabolism, protein folding and processing, type II secretory genes, aminophosphonate, methionine, and sulfur metabolism; butanoate metabolism; ion-coupled transporters; and other ABC transporter variants. The presence of these biased metabolic property distributions led to the conclusion that there is a preferential need for some metabolic processes in deep-sea adapted communities (DeLong *et al*, 2006). Years later Konstatinidis and colleagues created a larger sequence dataset from ~4000 m at HOT and used it for comparative analyses with other surface and deep metagenomes. The results from their metagenome comparison indicated that deep-sea microbes have acquired subtle molecular level adaptations to cope with the deep sea, and these adaptations include higher metabolic versatility to cope with the sparse and sporadic energy resources available represented by lager genome sizes, a preference for hydrophobic and smaller-volume amino acids in protein sequences and the absence of proteins found in surface-dwelling species, like UV repair enzymes or proteorhodopsin (Konstatinidis *et al*, 2009). The deep-sea community was also characterized by a larger average genome size and a higher content of transposases and prophages, whose

propagation is apparently favored by a more relaxed purifying (negative) selection in deeper waters (Konstatinidis *et al*, 2009).

Another study that was generated shortly after the first HOT metagenome, was the Mediterranean Sea 3000 m Km3 station metagenome. This metagenome was compared to the HOT station metagenomes and their metabolic profile was more similar to those of mesopelagic depth (500 – 700 m), rather than 4000 m sample. They suggested that in the absence of light, temperature is a major stratifying factor in the oceanic water column, overriding pressure at least over 4000 m deep (Martin-Cuadrado *et al*, 2007). In the Matapan-Vavilov Deep metagenome, also from the Mediterranean Sea but from a deeper site (4900 m) a comparative analysis of whole-metagenome data was performed. It revealed that unlike other deep-sea metagenomes, the prokaryotic diversity was extremely low, this possibly due to different environmental conditions such as high temperature, high salinity and high particulate and dissolved organic carbon (Smedile *et al*, 2013). In spite of its low diversity the Matapan-Vavilov metagenome possessed some deep-sea metabolic characteristics such as an abundance of *cox* genes. These encode different subunits of the carbon monoxide dehydrogenase (CoxL/CoxM/CoxS), and are also found in all other deep ocean metagenomes (Smedile *et al*, 2013). Carbon monoxide dehydrogenase is used to gain energy from the oxidation of carbon monoxide.

The Puerto Rico Trench (PRT) metagenome is thus far the only hadal metagenome. Analysis of its metabolic potential revealed an overabundance of genes associated with signal transduction mechanisms particularly of the PAS family of enzymes involved internal sensing of redox potential and oxygen, when compared to surface metagenomes. Genes encoding sulfatases for the degradation of complex

polysaccharides were also over represented. Inorganic ion transport and metabolism, along with transporters encoding outer membrane porins and genes involved in heavy metal efflux were also abundant (Eloe *et al,* 2011).

In the benthos a small number of deep-sea metagenomic studies have been performed using samples obtained from relatively shallow depths. Huang and colleagues studied sediment samples from the South China Sea by creating fosmid libraries from depths of 1,256 m, 1,330 m, 1,575 m and 2,893 m. Particular interest was drawn to a fragment identified as being most closely related to the *Gammaproteobacterium Idiomarina loihiensis*, and further analyses indicated that this microorganism may derive carbon and energy from the metabolism of tyrosine (Huang *et al*, 2009). More recently a sediment metagenome from the Sea of Marmara at 1250 m depth was generated, and genes involved in sulfate reduction, carbon monoxide oxidation, anammox and sulfatases were over-represented when compared to metagenome samples from surface and deep-sea seawater as well as soil and sediment samples (Quaiser *et al,* 2010*)*. At the whole-metagenome level the sediment sample was more similar to that of a soil metagenome than to either water column or subseafloor metagenomes.

In general metagenome studies deep-sea habitats have shown a high abundance of genes for functional categories of environmental sensing, signal transduction, transcription, transport and the use diverse carbon sources including recalcitrant organic compound and biopolymers, and in some cases alternative energy acquisition processes (Smedile *et al*, 2013; Eloe *et al*, 2011). Deep-sea microbial communities also possess subsystems and categories involved in heavy metal resistance and detoxification, such as

mercuric reductase and Co/Zn/Cd efflux system components, which may be indicative of particle associated lifestyles (Smedile *et al*, 2014).

In contrast with the microbial diversity in the surface ocean, mostly divided between proteobacteria (46%) and cyanobacteria (45%) (Ferreira *et al*, 2014), the composition of the microbial community appears to change with depth. Studies have shown that some groups of microbes are more abundant at depth than at the surface. Deep ocean communities in the water column are also dominated by *proteobacteria*, specifically *Alpha-* and *Gamma- proteobacteria* seem to dominate the 4000 m HOT site, 3000 m Km3 site and the PRT (Konstantinidis *et al,* 2009; Martín-Curado *et al,* 2007; Eloe *et al,* 2011). In the case of the metagenome from the Matapan-Vavilov Deep the microbial community was composed almost exclusively of *Gammaproteobacteria* (Smedile *et al*, 2013). In contrast with surface waters other groups of bacteria that seem to be over represented in the deeper ocean are SAR406, SAR202, *Planctomycetes* and *Gemmatimonadete* (Smedile *et al*, 2013; Eloe *et al*, 2011a,b). In deep-sea waters Thaumarchaeota also appear to also be a dominant contributor to the picoplankton community (Konstantinidis *et al*, 2009). However, these archaea are also found in the marine sedimentary subsurface; they penetrate several meters into the seafloor at organic-poor open ocean sites in the Equatorial Pacific (Teske, 2006) and in the Peru Basin (Sørensen *et al*, 2004).

In terms of deep-sea sediments the oxygenated surface contains relatively diverse bacterial communities in terms of numbers of phyla based on 16S rRNA gene clone libraries, from the *Alpha-*, *Delta-*, and *Gamma- proteobacteria*, *Acidobacteria*, *Actinobacteria*, and *Planctomycetes* (Orcutt *et al,* 2011). Organic-rich deep-sea sediments

also support bacterial communities dominated by the uncultivated OP9/JS1 phylum, whereas more organic-poor sediments host bacteria related to *Chloroflexi* and *Proteobacteria* (Inagaki *et al,* 2006, Kormas, *et al,* 2003, Webster *et al,* 2006). Many archaeal communities in organic-rich deep-sea sediments are dominated by the uncultivated deep-sea archaeal group/marine benthic group B (DSAG/MBGB) clade as well as the uncultivated miscellaneous crenarchaeotal group (MCG). These groups represent dominant archaeal lineages in clone libraries of archaeal 16S rRNA and occur in a wide range of sampling sites and sediment types (Inagaki *et al,* 2003).

While it is true that information about deep-sea metabolic processes is increasing due to the increasing number of metagenomic studies, it is also true that the PRT metagenome remains the only hadal metagenome to date. This implies that many of the patterns of adaptation in ultra-deep ocean environments are yet to be discovered.

The use of single cell genomes for understanding novel metabolic potential:

Another method for examining microbial identity and function at the genome level involves the use of single-cell isolation in conjunction with multiple displacement amplification (MDA) of DNA and genome sequencing (Raghunathan *et al*, 2005). This technology takes advantage of the isolation of single cells that can be acquired in a variety of methods, such as via the use of fluorescence activated flow cytometry, micromanipulation or microfluidics technologies (Lasken, 2012). MDA uses the unique properties of ϕ29 DNA polymerase (Blanco and Salas, 1984) and random primers to achieve > 1 billion fold amplification in a 30ºC isothermic reaction (Dean *et al*, 2001). The development of bioinformatics tools and genome assembly algorithms, targeting

specifically single cell genome assemblies, have provided the possibility to discern the desired DNA from potentially contaminants and low quality DNA sequences making, it feasible to study as yet uncultivated microbes (Bankevich *et al*, 2012; Lasken, 2012; Lasken and McLane, 2014). Given the sensitivity of the amplification process it is imperative that it occur in as sterile an environment as possible to avoid contamination. The value of single cell genomics has been exemplified by the numerous reports of single cell amplification of organism that belong candidate phyla (CP).

One of the first single cell genomes recovered from a candidate phyla was that of the CP TM7, which was isolated from a human host and also soil samples. (Marcy *et al*, 2007, Podar *et al, 2007*). The metabolic reconstruction of the TM7 single cell genomes provided the first look into their evolution and metabolic properties associated with their respective environments. For example, human associated TM7 genes were identified for type IV pilus biosynthesis, which has been suggested to facilitate adhesion to epithelial cells and may be involve in virulence (Marcy *et al*, 2007), whereas soil associated TM7 single cells were found to encode genes for plasmid acquisition, DNA repair , environmental stress responses and resistance to starvation, all of which may provide advantages for survival in soil environments (Podar *et al, 2007)*. Another early environmental single cell-derived partial genome was for a species of *Beggiatoa*, a marine group that inhabits sulfur-rich environments. It was predicted to perform sulfur oxidation but had not been successfully cultured preventing a physiological test of this hypothesis. Among the predicted genes encoded by the *Beggiatoa* were enzymes involved in sulfur oxidation, nitrate and oxygen respiration, and CO_2 fixation, confirming the chemolithoautotrophic physiology for these bacteria (Mußmann *et al, 2007)*.

More recently single cells were collected form nine different environments in an effort to understand the large diversity and genomic properties of microbes present in seawater, brackish water, freshwater and hydrothermal samples, among others (Rinke *et al*, 2013). This study provided in many cases the first recorded genomic information for many poorly described groups of organisms, including SAR406 (Marine Group A), OP3, OP8, WS3, and BRC1. Considerable insight was gained into the metabolic processes of these mostly uncharacterized organisms, among them novel amino acid use for the opal stop codon, the presence of archaeal-type purine synthesis in members of the Bacteria domain and the presence of bacterial-type sigma factors in members of the Archaea domain (Rinke *et al*, 2013). Based on the success of this relatively new technology it is clear that single cell genomics will continue to provide a genomic foundation for the understanding of many as yet uncultured novel microbial groups. This will include new revelations of the extent and nature of microbial diversity, its relationship to the environment, the discovery of the essential core or consensus genes of 'species' and the role of horizontal gene transfer in evolution (Lasken, 2007).

These techniques are especially suitable to deep trench environments because it makes it possible to investigate cells in a small sample, often the case with material obtained from ultradeep settings. Thus far the use of single cell genomics has provided information about some deep-sea microbes (Lloyd *et al,* 2013; Kaster *et al,* 2014). Eloe and colleagues, were able to amplify of 4 partial single cell genomes belonging to the *gamma-* and *alpha-proteobacteria*, *bacteriodetes* and *planctomycetes* from 6000 m in the PRT. The single cell metabolic profiles mirrored that of the findings in the PRT metagenome, including an abundance of metal efflux systems, TRAP and ABC

transporters, and sulfatases (Eloe *et al*, 2011). Although only a small percentage of each genome was recovered in the analysis, they provided an important glimpse into hadal associated metabolism.

Study sites:

The work presented in this dissertation focuses on two different ocean trenches, the Mariana Trench and the Puerto Rico Trench. The deepest surveyed location on Earth is the Challenger Deep within the Mariana Trench of the western Pacific Ocean that extends to depths at least as great as 10,971 m (Taira, *et al*. 2004; Taira, *et al*. 2005). This depth corresponds to about 111.2 megapascals ([MPa], 1,097.1 atmospheres, 16,123 pounds per square inch) of hydrostatic pressure. The Challenger Deep consists of three en echelon depressions along the trench axis, each of which is 6–10 km long, about 2 km wide, and deeper than 10,850 m. The eastern depression is believed to be the deepest, with a depth of 10,920 ± 5 m (Nakanishi and Hashimoto, 2010). Descents to the Challenger Deep have been accomplished sporadically since the middle of the twentieth century, including most recently during the Deep-Sea Challenge expedition led by James Cameron. Samples have been collected by deep trawls and bottom-grab sediment samplers (Beliaev, 1972; Mezhov, 1993; Quigley and Colwell, 1968), the highly publicized manned bathyscaphe Trieste (Piccard, 1960), free-falling/ascending vehicles (Mantyla and Reid, 1978; Yayanos, *et al*. 1981), deep CTD casts (Taira *et al*. 2005), deep current meter moorings (Taira *et al*. 2004) and the remotely operated vehicles Kaiko and Kaiko7000 (Kato *et al*, 1997; Nakajoh *et al*, 2007) and most recently the hybrid underwater robotic vehicle Nereus (Bowen *et al*, 2009). The most recent analysis of

organic carbon and associated microbial activity in the Mariana Trench showed that in comparison to a reference site at ~6000 m along the trench wall the cell counts in the Challenger deep at 10,900 m were consistently higher than at a shallower reference site through the first 20 cm of the sediments, accompanied by O_2 depletion, a sign of respiration (Glud *et al*, 2013). This is hypothesized to result from POM accumulation within the trench axis.

The Puerto Rico Trench (PRT) is the deepest location in the Atlantic Ocean at approximately 8,500 m, which translates to approximately 86.1 MPa (850 atmospheres and 12, 492 pounds per square inch) of hydrostatic pressure. It is a flat depression 1000 km long from east to west and is located between 50 to 100 km north of Puerto Rico (George and Higgins, 1979). Studies of the PRT have mostly focused on the geological characteristics of the trench, but some chemical and biological studies have been performed. It can be described as an oxygen rich oligotrophic environment with low flux of recalcitrant detrital material (Richardson *et al,* 1995). Low surface primary productivity of 0.4-2.2 mgC/m^2day translates into low levels of nutrients reaching the sediments (Couper, 1983). Surficial sediments (0-2 cm depth) from the PRT contain an average of 0.74% organic carbon (Richardson *et al,* 1995). The low percentage organic carbon are characteristic of older, more refractory organic material and contrast with the higher percentages of organic carbon present in younger, more labile material derived from shallow-water sources.

The studies that comprise this dissertation have as a main goal to better understand the metabolic capability and associated diversity of the hadal environments in the Puerto Rico Trench and Mariana Trench. It was possible to obtained microorganisms

from diverse and novel groups, in many cases poorly characterized groups and novel CP. This provided the opportunity to study novel microorganisms and their metabolic capabilities, including those less abundant members of the "rare biosphere" (Sogin *et al,* 2005). The research presented here is focused on the phylogenetic characterization and predicted metabolic properties of selected microbes and how that may translate into their role in the environment.

Chapter 2 presents analyses of single-cell genomes derived from microbes collected at 8,219 m within the PRT, to our knowledge at the time these were the deepest single cell genomes analyzed to date. The genomes included are those derived from cells belonging to Thaumarchaeota within the Archaea domain, and within the Bacteria domain SAR11 and two additional proteobacteria associated with previously cultured piezophilic organisms. The genomic properties of these deep-trench microbes are described with a focus on correlations with phylogeny and depth, and inferred physiological processes (e.g. carbon and energy acquisition). The results suggest that these microbes are all members of deep-sea clades. I found evidence for potential trench-specific adaptations, based on the unique metabolic properties encoded within the SAG genomes. These results illustrate new ecotype features that are likely to perform major roles in the adaptations of microorganisms to life at great depth.

Chapter 3 explores the expanded metabolic capability of candidate phylum OD1 using single cell genomics. Thirteen single cells were analyzed from sediment samples collected in the eastern depression of the Challenger Deep in the Mariana Trench as part of the Deepsea Challenge Expedition. The results indicate that members of the OD1 CP from the Challenger Deep have greater metabolic potential and structural complexity than

previously reported, including the finding of genes involved in oxidative phosphorylation, nitrate reduction and lipopolysaccharide synthesis.

Chapter 4 presents an analysis of six single amplified Marinimicrobia (SAR406) genomes (SAGs) from cells collected in the Challenger Deep within the Mariana Trench, during the Deepsea Challenge Expedition. These hadal microorganisms appear to take advantage of compounds such as carbon monoxide and hydrogen sulfide to supplement their energy requirements. Osmotic and oxidative stress regulation are also more abundant in the hadal Marinimicrobia than comparison genomes. Illumina-tag sequencing of bottom water samples collected in another region of the Challenger Deep reinforces the proposition that Marinimicrobia are abundant in this ecosystem

Lastly, Chapter 5 discusses the implications and future directions of the work presented this dissertation and how it has broadened scientific understanding of the microorganisms inhabiting the deepest parts of our planet's oceans.

REFERENCES

Allen E, Facciotti D, Bartlett D (1999). Monounsaturated but not polyunsaturated fatty acids are required for growth of the deep-sea bacterium Photobacterium profundum SS9 at high pressure and low temperature. *Applied and Environmental Microbiology* **65:** 1710-1720.

Allen E, Bartlett D (2002). Structure and regulation of the omega-3 polyunsaturated fatty acid synthase genes from the deep-sea bacterium Photobacterium profundum strain SS9. *Microbiology-Sgm* **148:** 1903-1913.

Aluwihare L, Repeta D, Chen R (2002). Chemical composition and cycling of dissolved organic matter in the Mid-Atlantic Bight. *Deep-Sea Research Part Ii-Topical Studies in Oceanography* **49:** 4421-4437.

Arakawa S, Sato T, Sato R, Zhang J, Gamo T, Tsunogai U, Hirota A, Yoshida Y, Usami R, Inagaki F, Kato C (2006). Molecular phylogenetic and chemical analyses of the microbial mats in deep-sea cold seep sediments at the northeastern Japan Sea. *Extremophiles* **10:** 311-319.

Arístegui J, Duarte CM, Gasol JM, Herndl GJ (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography* **54**.

Azam F, Malfatti F (2007). Microbial structuring of marine ecosystems. *Nature Reviews Microbiology* **5:** 782-791.

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A Lesin, VM, Nikolenko, SI, Pham, S, Prjibelski, AD, Pyshkin, AV, Sirotkin, AV, Vyahhi, N, Tesler, G, Alekseyev, MA, Pevzner, PA (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19:** 455-477.

Bartlett DH (2002). Pressure effects on in vivo microbial processes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1595:** 367-381.

Beja O, Koonin E, Aravind L, Taylor L, Seitz H, Stein J, Bensen DC, Feldman RA, Swanson RV, DeLong EF (2002). Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Applied and Environmental Microbiology* **68:** 335-345.

Beliaev GM (1972). Hadal bottom fauna of the world ocean.: Israel Program for Scientific Translations Ltd. p 157.

Blanco L, Salas M (1984). Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proceedings of the National*

Academy of Sciences **81:** 5325-5329.

Blankenship L, Yayanos A, Cadien D, Levin L (2006). Vertical zonation patterns of scavenging amphipods from the Hadal zone of the Tonga and Kermadec Trenches. *Deep-Sea Research Part I-Oceanographic Research Papers* **53:** 48-61.

Bowen AD, Yoerger DR, Taylor C, McCabe R, Howland J, Gomez-Ibanez D, Kinsey JC, Heintz M, McDonald G, Peters DB (2009) The Nereus Hybrid Underwater Robotic Vehicle for Global Ocean Science Operations to 11,000 m Depth. *Proc. Oceans*.

Campanaro S, De Pascale F, Telatin A, Schiavon R, Bartlett D, Valle G (2012). The transcriptional landscape of the deep-sea bacterium Photobacterium profundum in both a toxR mutant and its parental strain. *BMC Genomics* **13:** 567.

Certes A (1884). Sur la culture, à l'abri des germes atmosphériques, des eaux et des sédiments rapportés par les expéditions du Travailleur et du Talisman (1882–1883): *Compt. Rend. Acad. Sci.*

Chilukuri L, Bartlett D (1997). Isolation and characterization of the gene encoding single-stranded-DNA-binding protein (SSB) from four marine Shewanella strains that differ in their temperature and pressure optima for growth. *Microbiology-Uk* **143:** 1163-1174.

Couper AD (1983). *The Times Atlas of the Oceans [cartographic Material]*. New York; Toronto: Van Nostrand Reinhold Company.

DeLong E, Preston C, Mincer T, Rich V, Hallam S, Frigaard N, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496-503.

Edwards KJ, Becker K, Colwell F (2012). The Deep, Dark Energy Biosphere: Intraterrestrial Life on Earth. *Annual Review of Earth and Planetary Sciences* **40:** 551-568.

Eloe E, Fadrosh D, Novotny M, Allen L, Kim M, Lombardo M Yee-Greenbaum, J, Yooseph, S, Allen, EE, Lasken, R, Williamson, SJ, Bartlett, DH (2011a). Going Deeper: Metagenome of a Hadopelagic Microbial Community. *Plos One* **6**.

Ferreira AJ, Siam R, Setubal JC, Moustafa A, Sayed A, Chambergo FS, Dawe AS, Ghazy MA, Sharaf H, Ouf A (2014). Core microbial functional activities in ocean environments revealed by global metagenomic profiling analyses. *PloS one* **9:** e97338.

Fischer B (1894). Die Bakterien des Meeres nach den Untersuchungen de Plankton-Expedition unter gleichzeitiger Berüchksichtingung einiger alterer und neuerer

Untersuchunge. Ergeb. Plankton-Exped. pp 1–83.

France S (1993). Geographic-variation among 3 isolated populations of the hadal amphipod Hirondellea gigas (Crustacea, Amphipoda, Lysianassoidea). *Marine Ecology Progress Series* **92:** 277-287.

George RY, Higgins RP (1979). Eutrophic hadal benthic community in the Puerto Rico Trench. *Ambio Special Report***:** 51-58.

Gilbert JA, Dupont CL (2010). Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science* **3:** 347-371.

Glud RN, Wenzhofer F, Middelboe M, Oguri K, Turnewitsch R, Canfield DE, Kitazato, H (2013). High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature Geosci* **6:** 284-288.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology* **5:** R245-R249.

Herndl GJ, Reinthaler T, Teira E, van Aken H, Veth C, Pernthaler A Pernthaler, J (2005). Contribution of Archaea to Total Prokaryotic Production in the Deep Atlantic Ocean. *Applied and Environmental Microbiology* **71:** 2303-2309.

Huang Y, Lai X, He X, Cao L, Zeng Z, Zhang J, Zhou S (2009). Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in Escherichia coli. *Marine biotechnology* **11:** 124-131.

Inagaki F, Sakihama Y, Inoue A, Kato C, Horikoshi K (2002). Molecular phylogenetic analyses of reverse-transcribed bacterial rRNA obtained from deep-sea cold seep sediments. *Environmental Microbiology* **4:** 277-286.

Inagaki F, Suzuki M, Takai K, Oida H, Sakamoto T, Aoki K, Nealson KH, Horikoshi K (2003). Microbial communities associated with geological horizons in coastal subseafloor sediments from the Sea of Okhotsk. *Applied and Environmental Microbiology* **69:** 7224-7235.

Inagaki F, Nunoura T, Nakagawa S, Teske A, Lever M, Lauer A, Suzuki M, Takai K, Delwiche M, Colwell FS, Nealson KH, Horikoshi K, D'Hondt S, Jørgensen BB (2006). Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 2815-2820.

Jannasch H, Taylor C (1984). Deep-sea Microbiology. *Annual Review of Microbiology* **38:** 487-514.

Kaster A-K, Mayer-Blackwell K, Pasarelli B, Spormann AM (2014). Single cell genomic study of Dehalococcoidetes species from deep-sea sediments of the Peruvian Margin. *ISME J*.

Kato C, Li L, Tamaoka J, Horikoshi K (1997). Molecular analyses of the sediment of the 11000-m deep Mariana Trench. *Extremophiles* **1:** 117-123.

Kato C (2011). High Pressure and Prokaryotes. *Extremophiles Handbook*. Springer. pp 657-668.

Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station ALOHA in the North Pacific Subtropical Gyre. *Applied and Environmental Microbiology* **75:** 5345-5355.

Kormas K, Smith D, Edgcomb V, Teske A (2003). Molecular analysis of deep subsurface microbial communities in Nankai Trough sediments (ODP Leg 190, Site 1176). *Fems Microbiology Ecology* **45:** 115-125.

Lasken R (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology* **10:** 510-516.

Lasken R (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology* **10:** 631-640.

Lasken RS, McLean JS (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* **15:** 577-584.

Lauro FM, Chastain RA, Blankenship LE, Yayanos AA, Bartlett DH (2007). The Unique 16S rRNA Genes of Piezophiles Reflect both Phylogeny and Adaptation. *Applied and Environmental Microbiology* **73:** 838-845.

Le Bihan T, Rayner J, Roy MM, Spagnolo L (2013). <italic>Photobacterium profundum</italic> under Pressure: A MS-Based Label-Free Quantitative Proteomics Study. *PLoS ONE* **8:** e60897.

Lloyd K, Schreiber L, Petersen D, Kjeldsen K, Lever M, Steen A, Stepanauskas R, Richter M, Kleindienst S, Lenk S, Schramm A, Jorgensen BB (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496:** 215-+.

Marcy Y, Ouverney C, Bik E, Losekann T, Ivanova N, Martin H, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of*

America **104:** 11889-11894.

Martin-Cuadrado A, Lopez-Garcia P, Alba J, Moreira D, Monticelli L, Strittmatter A, Gottschalk G, Rodriguez-Valera F (2007). Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat. *Plos One* **2**.

Mezhov BV (1992). Two new species of the genus Macrostylis G.O. Sars, 1864 (Crustacea Isopoda Asellota Macrostylidae) from the Antarctic: Arthropoda Selecta. pp 83–87.

Mussmann M, Hu F, Richter M, de Beer D, Preisler A, Jorgensen B, Huntemann M, Glockner FO, Amann R, Koopman WJH, Lasken RS, Janto B, Hogg J, Stoodley P, Boissy R, Ehrlich GD (2007). Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *Plos Biology* **5:** 1923-1937.

Nakajoh H, Murashima T, Yamauchi N, Sezoko H (2007). Sediment Sampling at a Depth of 10,131m in the Challenger Deep by ROV Kaiko. *OCEANS 2007 - Europe***:** 1-6.

Nakanishi M, Hashimoto J (2011). A precise bathymetric map of the world's deepest seafloor, Challenger Deep in the Mariana Trench. *Marine Geophysical Research* **32:** 455-463.

Oger P, Jebbar M (2010). The many ways of coping with pressure. *Research in Microbiology* **161:** 799-809.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ (2011). Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiology and Molecular Biology Reviews* **75:** 361-422.

Pester M, Schleper C, Wagner M (2011). The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Current Opinion in Microbiology* **14:** 300-306.

Podar M, Abulencia C, Walcher M, Hutchison D, Zengler K, Garcia J, Holland T, Cotton D, Hauser L, Keller M (2007). Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology* **73:** 3205-3214.

Prieur D, Marteinsson VT (1998). Prokaryotes living under elevated hydrostatic pressure. *Biotechnology of Extremophiles*. Springer. pp 23-35.

Quaiser A, Zivanovic Y, Moreira D, López-García P (2010). Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *The ISME journal* **5:** 285-304.

Quigley MM, Colwell RR (1968). Properties of Bacteria Isolated from Deep-Sea Sediments. *Journal of Bacteriology* **95:** 211-220.

Raghunathan A, Ferguson H, Bornarth C, Song W, Driscoll M, Lasken R (2005). Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology* **71:** 3342-3347.

Reid JL, Mantyla AW (1978). On the Mid-Depth Circulation of the North Pacific Ocean. *Journal of Physical Oceanography* **8:** 946-951.

Reinthaler T, van Aken HM, Herndl GJ (2010). Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. *Deep Sea Research Part II: Topical Studies in Oceanography* **57:** 1572-1580.

Richardson MD, Briggs KB, Bowles FA, Tietjen JH (1995). A depauperate benthic assemblage from the nutrient-poor sediments of the Puerto Rico Trench. *Deep Sea Research Part I: Oceanographic Research Papers* **42:** 351-364.

Rinke C, Schwientek P, Sczyrba A, Ivanova N, Anderson I, Cheng J, Darling, A, Malfatti, S, Swan, BK, Gies, EA, Dodsworth, JA, Hedlund, BP, Tsiamis, G, Sievert, SM, Liu, WT, Eisen, JA, Hallam, SJ, Kyrpides, NC, Stepanauskas, R, Rubin, EM, Hugenholtz, P, Woyke, T (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Royer CA, Michael L. Johnson GKA (1995). Application of pressure to biochemical equilibria: The other thermodynamic variable. *Methods in Enzymology*. Academic Press. pp 357-377.

Simonato F, Campanaro S, Lauro F, Vezzi A, D'Angelo M, Vitulo N, Valle G, Bartlett DH (2006). Piezophilic adaptation: a genomic point of view. *Journal of Biotechnology* **126:** 11-25.

Smedile F, Messina E, La Cono V, Tsoy O, Monticelli L, Borghini M, Giuliano, L, Golyshin, PN, Mushegian, A, Yakimov, MM (2013). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environmental Microbiology* **15:** 167-182.

Sogin M, Morrison H, Huber J, Mark Welch D, Huse S, Neal P, Arrieta JM, Herndl GJ (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103:** 12115-12120.

Sorensen K, Lauer A, Teske A (2004). Archaeal phylotypes in a metal-rich and low-activity deep subsurface sediment of the Peru Basin, ODP Leg 201, Site 1231. *Geobiology* **2:** 151-161.

Stanley J, Konopka A (1985). Measurement of insitu activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology* **39:** 321-346.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland E, Dashiell P, Gomez ML,
Sieracki ME, DeLong EF, Herndl G, Stepanauskas R (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science* **333:** 1296-1300.

Taira K, Kitagawa S, Yamashiro T, Yanagimoto D (2004). Deep and bottom currents in the Challenger Deep, mariana trench, measured with super-deep current meters. *Journal of Oceanography* **60:** 919-926.

Taira K, Yanagimoto D, Kitagawa S (2005). Deep CTD casts in the Challenger Deep, Mariana Trench. *Journal of Oceanography* **61:** 447-454.

Teske A (2006). Microbial communities of deep marine subsurface sediments: Molecular and cultivation surveys. *Geomicrobiology Journal* **23:** 357-368.

Torsvik V, Øvreås L, Thingstad TF (2002). Prokaryotic Diversity--Magnitude, Dynamics, and Controlling Factors. *Science* **296:** 1064-1066.

Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro F, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi C, Bartlett DH,
Valle G (2005). Life at depth: Photobacterium profundum genome sequence and expression analysis. *Science* **307:** 1459-1461.

Vinogradova N (1997). Zoogeography of the abyssal and hadal zones. *Advances in Marine Biology, Vol 32* **32:** 325-387.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, Brochier-Armanet C, Chain PSG, Chan PP, Gollabgir A, Hemp J, Hügler M, Karr EA, Könneke M, Shin M, Lawton TJ, Lowe T, Martens-Habbena W, Sayavedra-Soto LA, Lang D, Sievert SM, Rosenzweig AC, Manning G, Stahl DA (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proceedings of the National Academy of Sciences* **107:** 8818-8823.

Webster G, John Parkes R, Cragg BA, Newberry CJ, Weightman AJ, Fry JC (2006). Prokaryotic community composition and biogeochemical processes in deep subseafloor sediments from the Peru Margin. *FEMS Microbiology Ecology* **58:** 65-85.

White D (2009). Modular Design of Li-Ion and Li-Polymer Batteries for Undersea Environments. *Marine Technology Society Journal* **43:** 115-122.

Yayanos A, Dietz A, Vanboctel R (1981). Obligately barophilic bacterium from the Mariana Trench. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **78:** 5212-5215.

Yayanos A (1995). Microbiololy to 10,5000 meters in the deep-sea. *Annual Review of Microbiology* **49:** 777-805.

Yayanos AA, Dietz AS, Van Boxtel R (1979). Isolation of a Deep-Sea Barophilic Bacterium and Some of Its Growth Characteristics. *Science* **205:** 808-810.

Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A, Ramette A (2011). Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLoS ONE* **6:** e24570.

ZoBell CE, Johnson FH (1949). The influence of hydrostatic pressure on the growth and viability of terestrials and marine bacteria. *Journal of Bacteriology* **57:** 179-189.

Chapter 2

Microbial Metabolic Properties at Greater Than 8,000 Meters Depth Within the Puerto Rico Trench Inferred From Single Cell Genomics

ABSTRACT

Hadal ecosystems occupy 45% of the total ocean depth within only 1-2% of the total area. However, the microbial communities in these ecosystems and their associated metabolic potential are largely uncharacterized. Here we present the analyses of four single amplified genomes (SAGs) obtained from 8,219 m within the hadal ecosystem of the Puerto Rico Trench (PRT): PRT *Nitrosopumilus*, PRT SAR11, PRT *Marinosulfonomonas,* and PRT *Psychromonas*. These microbes are all members of deep-sea clades, and two are closely related to previously isolated piezophilic (high pressure adapted) deep-sea microorganisms. The PRT *Nitrosopumilus* SAG, the only archaeal SAG examined, possesses genes associated with mixotrophy, including those associated with lipoylation and the glycine cleavage pathway, and may possess the ability to produce fatty acids and lipids, thought to be hallmark features distinguishing archaea from the other two domains of life. The PRT SAR11 SAG encodes for glycolytic enzymes previously reported to be missing in this highly abundant and cosmopolitan group. The PRT *Marinosulfonomonas and* PRT *Psychromonas* SAG possesses genes that may supplement its energy demands through nitrous oxide and hydrogen oxidation. We found evidence for potential trench-specific adaptations, as several SAG genes were observed only in a PRT metagenome and not in other shallower non-trench deep-sea metagenomes. These results illustrate new ecotype features that are likely to perform major roles in the adaptations of microorganisms to life at great depth.

INTRODUCTION

Little is known about the microbial communities in the deepest ocean environment, the hadal zone below 6,000 m. Most hadal zones are true trenches, lying within convergent margins wherein an oceanic plate is being subducted below a continental plate (Jamieson, 2011). Extreme environmental conditions, such as the lack of sunlight, recalcitrant organics, magma-crustal physical and chemical interactions, near freezing temperatures, and high pressure, give rise to unique ecosystems with distinct biological diversity. Hadal environments include specialized fauna that are distinct from their shallower deep-sea relatives (Blankenship *et al*, 2006; France, 1993).

The autochthonous microorganisms present and active in trenches and other deep-sea habitats are piezophilic, possessing optimal growth rates at pressures above atmospheric. Investigations of piezophiles indicate that high-pressure growth requires changes in membrane structure, DNA replication, protein synthesis, cell division and flagellar function; changes in pressure also lead to changes in the transcriptome, proteome and osmolyte levels (Bartlett, 2002; Lauro & Bartlett, 2008; Campanaro *et al*, 2012; Le Bihan *et al*, 2013). Piezophiles are also adapted for the utilization of complex organic carbon, for example, *Photobacterium profundum* SS9 is able to utilize chitin, cellulose, and pullulan (Vezzi *et al, 2005*).

Metagenomic studies of deep-ocean environments, 3-6 km depth, have provided the opportunity to examine the metabolic processes of as yet uncultivated microbes at a larger scale (Eloe *et al*, 2011b; DeLong *et al*, 2006; Martin-Cuadrado *et al*, 2007; Konstantinidis *et al*, 2009; Smedile *et al*, 2013). For examples, it has been observed that deep metagenomes contain an overabundance of *cox* genes, suggesting that many

piezophiles may be capable of using aerobic oxidation of CO as an additional energy source (Smedile *et al*, 2013). The Puerto Rico Trench (PRT) metagenome is thus far the only hadal metagenome. When compared to surface metagenomes the PRT has an overabundance of genes associated with signal transduction mechanisms particularly those encoding the PAS family of enzymes involved internal sensing of redox potential and oxygen, as well as genes encoding sulfatases for the degradation of complex polysaccharides. Inorganic ion transport and metabolism, along with transporters encoding outer membrane porins and genes involved in heavy metal efflux are also abundant in the PRT metagenome (Eloe *et al*, 2011b).

In deep-sea environments the absence of sunlight forces microorganisms to acquire carbon and energy from either exported production (White, 2009) or through chemoautotrophy. The extent of chemoautotrophy in ultra-deep trenches is not well understood and potentially limited (Eloe *et al*, 2011b). Beyond the photic zone at depths of ~800 m the uncultured groups ARCTIC96BD-19 and SAR324 appear to fix carbon by oxidizing reduced sulfur compounds (Swan *et al*, 2011). These two groups are also present in the PRT metagenome, suggesting this form of chemoautotrophy could be present in ultra-deep environments. Ammonia oxidizing archaea, mostly Marine Group I (MGI) Thaumarchaeota, are thought to play a major part in carbon fixation in the deep ocean and MGIs have also been reported to be prevalent in deep and ultra-deep environments (Francis *et al,* 2005; Eloe *et al*, 2011a). Within trench sediments chemoautrophic bacteria and archaea have been identified in Japan Trench cold seep communities (Arakawa *et al,* 2006; Inagaki *et al*, 2002).

In recent years the study of uncultivated microbial communities has motivated the

evolution of single-cell genomics techniques. The use of multiple displacement amplification (MDA) and optimized single cell genome assemblies has improved the available phylogenomic data and given insights into the metabolic capabilities of uncharacterized microorganisms (Lasken, 2012; Lasken & McLean, in press). Recently, single cell genomes were analyzed to address potential depth association of the SAR11 clade 1c, where genomes recruited more metagenomic sequence fragments from deeper environments than shallower ones (Thrash *et al,* 2014). Single cell genomics were used in the PRT by the amplification of 4 single cell genomes belonging to the *gamma-* and *alpha-proteobacteria*, *bacteriodetes* and *planctomycetes* (Eloe *et al*, 2011b). The data obtained was in agreement with findings in the PRT metagenome highlighted by the presence of metal efflux systems, different TRAP and ABC -like transporters, and different sulfatases (Eloe *et al*, 2011b). Although only a small percentage of each genome was recovered in the analysis, the described metabolic processes harbored within these single cells provided a look into hadal associated metabolism.

In this article analyses of single-cell genomes derived from microbes collected at a depth greater than 8,000 m within the PRT are presented, to our knowledge these are the deepest single cell genomes analyzed to date. The genomes included are those derived from cells belonging to MGI Thaumarchaeota, SAR11 and two other proteobacteria associated with previously cultured piezophilic organisms. The genomic properties of these deep-trench microbes are described with a focus on correlations with phylogeny and depth, and inferred physiological processes (e.g. carbon and energy acquisition).

MATERIALS AND METHODS

Collection and sorting

Seawater and amphipods were collected using a free falling vehicle (FFV; Figure S2.3) deployed during November 2010 on board the Makai (50 ft catamaran) over a water column depth of 8,219 m (+/- 66 m) within the PRT (19º 46.022' N, 66º 55.432' W). Microbial samples were collected with a pair of baited 30 l Niskin water sampling bottles. Recovered seawater and amphipods were placed inside polyethylene bags, pressurized to 62 megapascals (MPa, 9,000 pounds per square inch) and held at a temperature of 4ºC. Samples were transferred to the J. Craig Venter Institute (JCVI) for single-cell sorting. The collected amphipod was homogenized using an autoclaved pestle in a microcentrifuge tube. The sample was filtered and stained with 10x SYBR Green I fluorescent dye (Invitrogen, Carlsbad, CA) and then sorted using a cooled FACS-Aria II flow cytometer (Becton Dickenson and Company, Franklin Lakes, New Jersey) and stored at –80ºC for later processing.

Genome amplification and sequencing

Genomic material was amplified using multiple displacement amplification (MDA) in a 384-well format using a GenomiPhi kit (GE Healthcare, Waukesha, WI) and a custom BioCel robotic system (Agilent Technologies, Santa Clara, CA) as described by McLean et al (2013). 16S rRNA genes were PCR amplified, cleaned and amplicons were sent for Sanger sequencing at the Joint Technology Center (JTC, J. Craig Venter Institute, Rockville, MD). Resulting 16S rDNA sequences were evaluated for evidence of contaminated sequences and those were removed from consideration for whole genome sequencing. 16S rDNA sequences were compared to the NCBI nr/nt database using

BLASTN (Altschul *et al*, 1990) and organisms of interest were selected based on phylogenetic novelty. DNA recovered from 40 cells was prepared for whole genome sequencing via the Illumina HiSeq 2000 platform. For comparison and validation, an MDA amplified *Escherichia coli* (*E. coli*) sample was processed along with all other selected genomes (data not shown). Libraries were prepared using the multiple barcode technology of the Nextera™ DNA Sample Prep Kit (Illumina, San Diego, CA) and sent to JCT for sequencing.

Assembly, annotation and genome completion

Sequences were assembled using the Spades assembler, SPAdes 2.3.0 (Bankevich *et al*, 2012). Four genomes were selected for further processing and annotation. Assembled genomes were uploaded to the IMG-ER platform (https://img.jgi.doe.gov/cgi-bin/er/main.cgi, Markowitz *et al*, 2014) for genome annotation.

16S rRNA gene sequences recovered from each SAG were analyzed by BLASTN against the NCBI nr/nt database (Altschul *et al*, 1990). Sequence matches that were at a minimum of 95% similarity and 97% alignment were extracted and used for phylogenetic reconstruction, in some cases other phylogenetically relevant species were added as references. All reference sequences extracted from NCBI were also annotated for associated environmental source, including water column depth, if available. Sequences were aligned with the SINA aligner (http://www.arb-silva.de/aligner/, Pruesse *et al*, 2012) and maximum-likelihood tree were created using FastTree (Price *et al*, 2009) and RaxML within the CIPRES Science Gateway V.3.3 (https://www.phylo.org/)

Genome-encoded protein predictions were obtained from IMG-ER and classified phylogenomically using DarkHorse software, version 1.4 (http://darkhorse.ucsd.edu/, Podell & Gaasterland, 2007). DarkHorse results were used to identify potential contaminating sequences among SAG contigs, based on whether or not taxonomic lineages associated with predicted proteins on each assembled contig were similar to or different from the rest of the contigs (Jones et al 2011). Estimated genome completeness was calculated using the Human Microbiome Project protocol for bacterial genomes (http://hmpdacc.org/tools_protocols/tools_protocols.php) and the protocol of Podell et al. (2013) for the archaeal genome. Genome size was estimated based on the most closely related previously sequenced microbe.

SAG comparisons to known genomes and metagenomes

Functional comparisons were performed using the IMG-ER platform. Using each SAG's most closely related genomes (MCGR), side-by-side comparisons of COG categories, pfams and KEGG pathways were performed. In addition SAGs and their MCGR were evaluated using FR-HIT read recruitment software with default parameters (Niu *et al*, 2011), against 96 metagenomic sets from the Global Ocean Survey (GOS; (Venter *et al*, 2004; Rusch *et al*, 2007; Brown *et al*, 2012) and metagenomic samples from the Hawaii Ocean Time series (HOT; DeLong *et al*, 2006; Konstantinidis *et al*, 2009) and Mediterranean Sea (Martin-Cuadrado *et al*, 2007, Smedile *et al*, 2013) (Table S2.4). Recruitments were normalized based on the genome size of the analyzed genome and total number of reads of the analyzed metagenome. To obtain a more detailed representation of the PRT SAG genes in the metagenomes, best reciprocal blasts (BRBs)

were performed. Best reciprocal blast analyses were generated by extracting sequences from FR-HIT results that contained more than 60 bp in alignment and 75% sequence similarity for each specific SAG and its comparison genomes. Recovered sequences were compared using BLASTN to small databases generated for each SAG and its comparison genomes. RBBs hits to each of these genomes were quantified and normalized as described above. Reads that preferentially matched the SAG genomes were considered true read recruitments, while reads that preferentially matched any of the other comparison genomes were excluded from the analysis.

RESULTS

Phylogenetic placement and biogeography of MDA amplified single cells

A total of 15,720 single-cells were sorted from undiluted trench seawater and amphipod associated microbial communities. Amphipods belonged to the genus *Hirondellea* (Carvajal & Rouse, unpublished results). Subsequently, the DNA from 2,880 of these sorted cells was amplified by MDA, with 22% of the amplified DNA yielding positive amplification.

The 16S rRNA gene sequences obtained from the MDA amplified single cells included a large number of microbial phyla and classes, among them *Alpha, Beta* and *Gamma proteobacteria, Bacteriodetes, Fusobacteria, Firmicutes, Actinobacteria and Acidobacteria* (Figure S2.4). Forty single-cells were selected based on phylogenetic novelty for whole genome amplification. Of these, the four containing the greatest genome completeness were selected for further annotation and analysis: PRT *Nitrosopumilus,* PRT *Marinosulfonomonas,* PRT SAR11 and PRT *Psychromonas* (Figure

2.1). PRT *Nitrosopumilus,* falls within the I.1a group of the Thaumarchaeota, the same group as *Nitrosopumilus maritimus* (Könneke *et al*, 2005; Pester *et al*, 2011). Most of the members of this clade were recovered from sediment and water columns samples at depths ranging from 3,500 m to 5,000 m below the sea surface (Durbin & Teske, 2010) (Figure 2.1).

Most of the PRT *Marinosulfonomonas* SAG-related sequences align within a major clade of primarily uncultivated environmental samples isolated from deep trenches, hydrothermal vent plums, and *Riftia* tube worms (Hügler *et al*, 2010). PRT *Marinosulfonomonas* is a member of the *Alphaproteobacteria* class clustered in a monophyletic clade within the alpha-3 subgroup of the *Roseobacters*. PRT *Marinosulfonomonas* is closely related to *Rhodobacterales* PRT1, the first piezophilic microorganism cultured from the PRT (Figure 2.1; Eloe *et al*, 2011c). PRT1 and PRT *Marinosulfonomonas* belong to a phylogenetically distinct clade dominated by deep ocean bacteria within the *Roseobacter* lineage of the *Rhodobacterales*. Other closely related sequences included microbes inhabiting low temperature Arctic and Antarctic marine environments.

Previous studies have suggested that deep-sea SAR11 (770 m) fall within the SAR11 group 1c (Thrash *et al*, 2014). In contrast, PRT SAR11 SAG falls within the SAR11 group II, a sister clade to the SAR11 group I (1a, b and c) which includes the genus *Pelagibacter*. SAR11 group II is divided into two distinct clusters (Vergin *et al*, 2013), with SAR11 SAG falling within a subclade of the group IIa, which is composed predominantly of sequences derived from deeper marine environments ranging from 2,400 to 3,300 meters below the sea surface (Shaw *et al*, 2008) (Figure 2.1).

PRT *Psychromonas* SAG falls within the genus *Psychromonas* within the *Gammaproteobacteria*. PRT *Psychromonas* SAG clusters together with the cultivated piezophile *Psychromonas* sp. strain CNPT3 (CNPT3) and other microbes belonging to deep-ocean environments (e.g. trenches and whale falls; Goffredi *et al*, 2005) (Figure 2.1). CNPT3 was the first piezophile ever isolated, and it is known to possess various adaptations for growth at high pressure (DeLong *et al*, 1997, Lauro *et al*, 2013). Members of this clade can clearly be assigned to deep-sea habitats.

Genomic characterization

Assembled sequence information ranged from 0.6 Mbp to 1.7 Mbp per genome. Genome completion varied from 59% to 77% completion (Table 2.1). Each genome was assigned a "most closely related genome" (MCRG) based on predicted protein homology acquired from the DarkHorse analysis (Podell & Gaasterland, 2007). Gene count, number of contigs, GC%, COG count and coding region % was obtained from the IMG platform. The estimated genome size was calculated based on the % completion and the size of the MCRG (Table 2.1). All of the calculated genome sizes were smaller than their MCRGs and they all have a larger percentage of non-coding space than their associated MCRGs (Table 2.1).

The large fraction of non-coding regions in the PRT SAGs may be partially explained by sRNAs (Table S2.6). The PRT *Psychromonas* possesses sRNAs that are most closely related to those present in CNPT3, for example protein binding sRNAs 6S and RNase P. The PRT SAR11 encodes a transfer-messenger RNA (tmRNA) that is most similar to a sequence collected from 4000 m at ALOHA station. PRT

Marinosulfonomonas also encodes a tmRNA closely related to a tmRNA found in members of the *Rhodobacterales*, a C4 antisense RNA associated with bacteriophages and number of sRNA annotated as pseudoknots, which form secondary structures associated with viral translation (Wang *et al*, 1995).

General metabolic comparisons

To assess whether the PRT SAGs encode unique metabolic characteristics, proteins were assigned COG, KEGG and Pfam categories and compared between the SAGs and their associated MCRGs (Table S2.7). For the sake of brevity, COG categories alone are described below.

PRT *Nitrosopumilus* SAG shares 358 out of 406 of its annotated COGs with the thaumarchaeaon *Nitrosopumilus maritimus* SCM1. Among the shared COGs are several carbon fixation genes involved in the 3-hydroxypropionate-4-hydroxybutyrate pathway and general metabolic genes involved in cell functions. 48 COGs unique to the SAG were distributed among 13 different categories. Approximately 30% of the 48 unique COGs were unknown (catergory [S]), followed by amino acid transport and metabolism (E;20%) and then [C] 7% and [O] 7% (Figure S2.6). The *PRT Nitrosopumilus SAG* can be differentiated from *N. maritimus* by the presence of unique sequences that may provide a selective advantage in the deep ocean, including enzymes for urea degradation, lipoic acid synthesis, glycine cleavage and remarkably, fatty acid synthesis (Table 2.2).

The PRT *Marinosulfonomonas* SAG is related to *Alphaproteobacterium* members of the *Rhodobacterales* family. These microorganisms exhibit extensive metabolic diversity, a characteristic that is also reflected in the SAG genome. The PRT

Marinosulfonomonas SAG shares 783 out of 913 COGs with the *Alphaproteobacterium Thalassiobium* R2A62. Among their shared metabolic properties are housekeeping functions associated with tRNA synthases, pilus synthesis and assembly, cellular shape, transport systems and metabolic processes. 130 COGs are unique to PRT *Marinosulfonomonas* SAG. These are distributed among the 19 different categories, with the most abundant belonging to function unknown [S] (39%) and general function prediction [R] (15%) (Figure S2.6).

PRT SAR11 SAG shares 496 out of its 536 annotated COGs with *Candidatus Pelagibacter ubique* SAR11 HTCC1062 (*P. ubique*). The forty COGs unique to this SAG are distributed among 12 different categories (Figure S2.6). These include novel metabolic properties not present in other described SAR11 members, among them genes that encode for the enzymes phosphofructokinase-6 and pyruvate synthase.

The PRT *Psychromonas* SAG was compared with its MCRG, the cultured piezophilic bacterium *Psychromonas sp*. strain CNPT3 (CNPT3), as well as with the surface water bacterium *Psychromonas ingrahamii*. PRT *Psychromonas* SAG is found to share a number of genes with CNPT3 that are not found in the shallow-water *P. ingrahamii*. This includes genes involved in motility, and a number of transporters and permeases for iron, multidrug and sugar and amino acid transport. These two microbes also share the ability to produce a periplasmic nitrate reductase system protein NapA and encode the carbon starvation protein CstA, which is suggested to be involved in peptide transport under stressed conditions (Rasmussen *et al*, 2013). PRT *Psychromonas* SAG shares 559 out of its 663 COGS with *P. ingrahamii* and 632 with CNPT3. Seventy four

COGs are unique to the SAG when compared to *P. ingrahamii* and thirty one when compared to CNPT3 (Figure S2.6).

Compared to *P. ingrahamii* the genome of CNPT3 is also enriched in type B fatty acid synthase (FAS)-polyketide synthase (PKS) genes for the production of polyunsaturated fatty acids (PUFAs; Lauro *et al*, 2013), PUFAs perform important roles in low temperature and high-pressure adaptation (Usui *et al*, 2012). It is not known whether PRT *Psychromonas* also contains these genes. Novel metabolic characteristic are described for each (Table 2.2) and further discussed below.

DISCUSSION

Genomic characterization

Calculations of the genome sizes of cultured piezophiles, and of the genome sizes inferred from metagenomic and single-cells genomic analyses, have led to the conclusion that deep-sea microbes have larger genomes than their surface relatives (Lauro & Bartlett, 2008; Eloe *et al*, 2011b; Thrash *et al*, 2014; DeLong *et al*, 2006; Konstantinidis *et al*, 2009). This has been ascribed to the reduction in purifying selection encountered in deep-sea versus shallow-water habitats (Konstantinidis *et al*, 2009). However, all four of the SAGs described here possess smaller genome sizes than their MCRGs. The reason for this difference is unknown but could stem from environmental characteristics such as nutrient regime, depth, or benthic boundary layer association. Smaller genome sizes could stem from methodological differences in the calculations used or gene set analyzed. Certainly the four genomes examined in this study are too small in number to be a statistically significant data set. SAGs from natural bacterioplankton often have reduced

sizes, an observation that has been interpreted as reflecting genome streamlining and adaptation to oligotrophic environments (Swan *et al,* 2013).

Even with the reduced genome size, most of the PRT SAGs have a larger percentage of non-coding regions than their associated MCRGs (Table 2.1). Piezophilic microbes contain larger than average intergenic regions (>150 bp) (Lauro & Bartlett, 2008), and the piezophilic bacterium *Photobacterium profundum* expresses at high pressure a large number of small transcripts encoded within intergenic regions, many of which appear to be cis-acting regulatory RNAs (Campanaro *et al*, 2012). Intergenic regions are important for controlling many metabolic processes via transcriptional and post-transcriptional regulation. In this regard it is of note that transfer-messenger RNA sequences were found in the PRT SAR11 and PRT *Marinosulfonomonas* SAG. Transfer-messenger RNAs are involved in the degradation of incompletely synthesized peptides from truncated mRNA, as well as the recycling of stalled ribosomes through the *trans*-translation system. This occurs primarily under stressed conditions (Muto *et al*, 2000), and could represent a useful adaptation to the extreme environmental conditions of the ultra-deep ocean.

Novel metabolic potential

Lipoylation /Glycine Cleavage/Ammonia Acquisition

Members of the division *Thaumarchaea* are among the most abundant archaea on the planet (Pester *et al*, 2011). Characterized by their ability to oxidize ammonia autotrophically, members of the *Thaumarchea* have been suggested to play a major role in the nitrogen cycle, particularly in the deep ocean (Konstantinidis *et al*, 2009; Herndl *et*

al, 2005). PRT *Nitrosopumilus* is the deepest member of *Thaumarcheae* to be studied at the genomic level in detail. The identification of genes (Table 2.2) associated with lipoylation, glycine cleavage system (GCS), fatty acid metabolism, and Lipid A biosynthesis implies that this archaeon contains fatty acids, a property not yet demonstrated in any archaeon. Some of the unique PRT *Nitrosopumilus* sequences also indicate that the deep members of this clade possess multiple strategies for ammonia acquisition.

Lipoate or lipoic acid (LA) is a highly conserved cofactor in the aerobic metabolism of 2-oxoacids and C1 compounds (Posner *et al*, 2013). In *E. coli,* where lipoylation has been extensively studied, 3 enzymes, LlpA, LipA, and LipB, carry out two different lipolation reactions to the same end (Morris *et al*, 1994). These eznymes catalyze attachment of the lipoyl moiety to dihydrolipoyl acyltransferase (E2), to form lipoyl cofactor, which is required for the function of several key enzyme complexes in oxidative and one-carbon metabolism. When LA is available, organisms generally use the LplA pathway, but if not, they will synthesize LA de novo from octanoic acid using LipA and LipB. PRT *Nitrosopumilus* SAG is the first known member of the Thaumarchaea to possess lipoylation as a part of its metabolism.

It is important to note that E2 is missing from the incomplete genome of PRT *Nitrosopumilus* SAG. In most *Thermococcus* species (one of the two archaeal genera where lipoylation has been studied) E2 is also absent from their genomes. However, in these systems the glycine cleavage system protein H is used as a lipoylation target (Borziak *et al*, 2014). The PRT *Nitrosopumilus* SAG has all the genes for a complete glycine cleavage system, suggesting that protein H is likewise its target for lipoylation.

Within PRT *Nitrosopumilus* the glycine cleavage system may also be used to acquire ammonia. The catabolism of glycine involves a reversible reaction whereby glycine is cleaved to carbon dioxide, ammonia and a methylene group (-CH_2-), which are each used in subsequent catabolic reactions. The methylene group is accepted by tetrahydrofolate (THF) to form 5,10-methylene-THF. The 5,10-methylene-THF molecule is involved in purine and methionine biosynthesis. The regeneration of THF produces NADH, which can be used directly to yield energy, and ammonia, which can be utilized for a variety of processes, including energy generation via ammonia oxidation. GCS has only been studied in hyperthermophilic and halophilic archaea (Fischer *et al*, 2012, Lokanath *et al*, 2004).

It has been also suggested that ammonia-oxidizing Thaumarchaea utilize ureases to catalyze the degradation of urea to carbon dioxide and ammonia when environmental ammonia concentrations are low (Lu & Jia, 2012). PRT *Nitrosopumulas* encodes all components of the urease enzyme as well as urease accessory proteins. In addition, *Nitrosopumulas* SAG also possesses a GlnK gene, which is a well-known regulator of ammonium transport and incorporation in Eukarya, Bacteria and selected groups of Archaea (within the Euryarchaeota; Leigh & Dodsworth, 2007).

Fatty Acid Metabolism and Lipid Synthesis

The *Nitrosopumilus* SAG encodes for a number of genes associated with fatty acid and lipid synthesis, including 3-oxoacyl-[acyl-carrier-protein] synthase III (KAS III), which is important for the production of monounsaturated fatty acids (Lai & Cronan, 2003) required for adaptation to high pressure and cold temperature (Allen & Bartlett,

2002). No other archaeal KASIII enzyme has been reported previously (Lombardo *et al*, 2012). When compared to all public genomes available in IMG, the top hit was to a hypothetical protein from a single cell genome sequence of an environmental Thaumarchaeota archaeon (SCGC AAA282-K18, unpublished results). The *Nitrosopumilus* SAG also encodes for an acetyl-CoA carboxylase, which catalyzes the conversion of acetyl-CoA to malonyl-CoA, providing the substrates needed (acetyl-CoA and malonyl-CoA) for KASIII to perform the first condensation step in fatty acid synthesis.

In addition to the *Nitrosopumilus* SAG, KASIII is also encoded in the SAR11 SAG, despite not being present in any other *Pelagibacter* like organisms to date. The top BLASTP score match shows 38% (at 99% coverage) similarity with *Mariprofundus ferrooxydans*, a deep-sea iron oxidizing microbe (Singer *et al*, 2011).

The *Nitrosopumilus* SAG contained an acetyltransferase involved in Lipid A synthesis (Table 2.2). Given that archaeal cells are not known to produce lipopolysaccharide as a part of their cell membrane it is difficult to speculate on the significance of finding enzymes belonging to the Lipid A biosynthetic pathway or to the potential role of Lipid A. In *E. coli,* Lipid A can undergo acylation with palmitoleate instead of laurate, and although the reason for this adaptation is not clear, it has been suggested that this alteration might function to adjust outer membrane fluidity in *E. coli* cells shifted to low temperatures (Carty *et al*, 1999).

Carbon and energy acquisition

Some organisms, including those in low-temperature and high-pressure environments, have the ability to reduce nitrous oxide (N_2O) to nitrogen gas (N_2) without performing the complete denitrification pathway (Sanford *et al,* 2012). Thus, since PRT *Marinosulfonomona* contains genes associated with N_2O reduction it is possible that it employs N_2O in the oxidation of organic matter. It is possible that PRT *Marinosulfonomonas*, which was isolated from a sediment-seawater interface, makes use of the N_2O produced from incomplete denitrification processes present in sediments. It has also been proposed that nitrification leads to N_2O production in sinking particles (Wilson *et al*, 2014), which may represent microenvironments for these organisms in the deep sea.

PRT *Marinosulfonomonas* encodes for a collagenase, and related proteins are also present in other members of the *Rhodobacterales*. These enzymes may be needed to take full advantage of complex polymers associated with sinking particles or sedimentary substrates.

Key genes involved in glycolysis are found in PRT SAR11 SAG, which provides insight into its carbon utilization (Table 2.2). The only other described SAR11-like genome that possesses pyruvate kinase and phosphofructokinase enzymes is the *Alphaproteobacterium* HIMB59 (Grote *et al*, 2012). However, the phylogenetic affiliation of HIMB59 is not fully resolved (Viklund *et al*, 2013). More detailed studies of these two enzymes must be done to clearly understand their evolution and distribution within the SAR11. Nevertheless, the presence of these enzymes in the PRT SAR11 SAG together with the numerous ABC-type sugar transporters also present suggests that this

microorganism may be capable of glycolysis via the Embden–Meyerhof–Parnas pathway, utilizing sugar substrates for carbon acquisition and energy production. In addition, three enzymes that catalyze the first three reactions in the myo-inositol degradation pathway that ultimately feeds into glycolysis are also present in this SAG.

Analyzing the metabolic potential of PRT *Psychromonas* SAG provided a unique opportunity to address functions shared among related piezophiles present in different ocean basins. Genes associated with carbon acquisition and energy generation are present in the PRT *Psychromonas* SAG and its deep-sea MCRG but not in its surface MCRG. Periplasmic nitrate reductase (NapBDEF) initiates aerobic ammonification and is thought to be involved in the disposal of excess reductant power and as an electron sink to regenerate NAD^+, aiding carbon acquisition and cell growth processes (Richardson, 2000). The comparison between these two deep-sea adapted organisms and their surface water relatives suggest that the deep-sea psychromonads have evolved the capability to utilize a greater variety of organic matter.

Among the unique metabolic properties of the PRT *Psychromonas* SAG is a gene that codes for a Ni,Fe-hydrogenase I enzyme. This is a membrane bound protein that links H_2 oxidation to anaerobic or aerobic respiration, with recovery of energy via protonmotive force (Vignais & Billoud, 2007). Ni,Fe-hydrogenases have been studied extensively in organisms associated with hydrothermal vents and anaerobic systems, but less is known from microorganisms inhabiting oxygenated marine environments (Kim *et al,* 2011). It is important to note that CNPT3 encodes a different kind of Ni,Fe-hydrogenase that belongs to the group 4 of hydrogenases. Enzymes in this group reduce protons from water to dispose of excess reducing equivalents produced by the anaerobic

oxidation of C1 organic compounds such as carbon monoxide or formate (Vignais & Billoud, 2007).

Sulfur Metabolism

Some microorganisms, including some members of the SAR11, lack the ability to utilize sulfate and must acquire their sulfur using different sulfur-containing substrates (Tripp *et al*, 2008). Two of the PRT SAGs, PRT SAR11 and PRT *Marinosulfonomonas*, encode for a taurine dioxygenase (TauD) gene. In *E. coli* TauD is used to provide an alternate sulfur source under sulfur deficient conditions (Eichhorn *et al*, 1997). The taurine degradation pathway produces sulfite for cysteine biosynthesis. It is possible that PRT *Marinosulfonomonas* SAG and PRT SAR11 SAG utilizes taurine as their preferred source of reduced sulfur.

Osmoregulation

Genes encoding for proteins associated with osmotic regulation are also present in several of the SAGs. For example, the NhaD type sodium/proton-antiporters, which have been proposed to serve osmoregulatory purposes (Kurz *et al*, 2006) are found in the *Nitrosopumilus* SAG. Furthermore, the *Nitrosopumilus* and *Marinosulfonomonas* SAGs encode for aquaporins. Aquaporins are known to be important in osmotic pressure adaptation by effluxing water from cells exposed to hypotonic environments. Aquaporins have also been suggested to be especially useful for the retention of small molecule, compatible solutes like urea, glycerol, and glucose (Kumar *et al*, 2007). Aquaporin-4, encoded in the PRT *Nitrosopumilus SAG* has only previously been described as a

mammalian protein, while the *Marinosulfonomonas* SAG encodes an aquaporin Z (AqpZ). The role of AqpZ in free-living marine microorganisms has not been fully characterized. However, it has been discovered and characterized in *E. coli* as a channel for rapid water efflux across the membrane, helping microorganisms to cope with osmotic downshift (Calamita 2000). This channel is selectively permeable to water, has a role in both the short-term and the long-term osmoregulatory response, and is required by rapidly growing cells. Given that osmotic pressure and hydrostatic pressure can have opposing effects on macromolecules (Robinson & Sligar, 1995), and deep-ocean organisms accumulate large amount of osmolytes (Yancey *et al*, 2014) (sometimes referred to as piezolytes; Martin *et al*, 2002) aquaporins could play a role in high-pressure adaptation.

Transporters

The ability to transport and use complex organic compounds may provide an advantage over microorganisms with narrower organic substrate catabolic processes in the ultra-deep ocean. All of the PRT SAGs encoded diverse transporters, from ABC – sugar/multidrug or peptide transporters to heavy metal efflux proteins (Table 2.2). The ability of these microbes to intake a variety of compounds maybe an adaptation to dealing with an environment that has sporadic and varied nutrient availability.

Best Reciprocal Blasts (BRB)

Phylogenetic analyses demonstrated that the PRT SAGs were most closely related to other deep ocean microorganisms. BRB analysis also indicated that the PRT

Nitrosopumilus and PRT SAR11 SAGs recruited preferentially to deep ocean metagenomes in comparison to surface metagenomes (Figure 2.2). This is not apparent for the other two SAGs, which may be due to the dramatically lower abundance of all BRB hits for all metagenomes examined when these SAGs and their comparison genomes were examined (Figure S2.7).

Several SAG genes were found in all metagenomes, but no SAG genes were generally unique among deep ocean metagenomes (PRT, HOT 4000 m and Deep Mediterranean 3000 m). However, these read recruitments did identify genes uniquely recruiting to the PRT metagenome (Table 2.3). Among the genes uniquely represented, 30% are related to known genes involved in transcriptional regulation and 15% are related to genes associated with transporters. Of these unique genes, no one gene or set of genes were found across all PRT SAGs. The lack of a conserved "deep gene" necessary for adaptation to deep ocean conditions could stem from the under sampling of ultra-deep ocean environments.

CONCLUSION

The objective of this study was to assess the diversity and metabolic capabilities of microbes present in the deepest part of the Atlantic Ocean within the PRT. The use of single-cell genomics enabled the amplification and sequencing of four partial genomes. Two of the SAGs are of special interest (PRT SAR11 SAG and PRT *Nitrosopumilus* SAG) as they are the deepest studied members of highly abundant groups of large-scale biogeochemical significance (Morris *et al*, 2002; Schattenhofer *et al*, 2009; Karner *et al*,

2001; Teira *et al*, 2006). Phylogenetic analyses of all four SAGs indicate that they are autochthonous residents of deep-ocean environments.

Genes present in the SAGs but absent in their comparison MCRGs revealed novel metabolic capabilities including those associated with nitrogen, sulfur, carbon, and energy acquisition mechanisms. Among some of the most significant findings presented are the potential for MC1 Thaumarchaea to synthesize fatty acids and the ability for PRT SAR11 to perform complete glycolysis. Also significant when considering survival in the PRT are genes involved in generating energy from H_2 or N_2O oxidation by PRT *Psychromonas and* PRT *Marinosulfonomonas* SAGs respectively. The importance of osmoregulation in the ultra-deep ocean is suggested by the finding of aquaporins in PRT *Marinosulfonomonas and* PRT *Nitrosopumilus*.

More single cell genomic, metagenomic and environmental systems biology studies that target the hadal condition will be needed to better highlight the evolutionary, genetic and regulatory changes required for bacterial and archaeal life in the deepest portions of the world's ocean.

Nucleotide sequence accession number

These single cell genomes have been deposited at DDBJ/EMBL/GenBank under the accession numbers JPUE00000000, JPUP00000000, JPUQ00000000 and JPUR00000000. The annotated genomes are also available in the IMG-ER platform, IMG-Genome-IDs: 2518645502, 2518645503, 2518645501 and 2518645504.

ACKNOWLEDGEMENTS

We would like to thank Jessica Blanton, Juan Ugalde, Greg Rouse, Carlos Rios-Velazquez, Logan Peoples, Graham Wilhelm, Shelbi Randenberg, Jessica Wdowiarz and Justin DeShields for their contribution to discussion and collection of samples and data. We are also grateful for the financial support provided by the National Science Foundation grants 0801973 and 0827051 and National Science Foundation Graduate Research Fellowship 068775, the National Aeronautics and Space Administration (NNX11AG10G), the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number T32GM067550, the UCSD Academic Senate, the Department of Scripps Institution of Oceanography and a generous contribution from Joanie Nasher.

Chapter 2 is a full-length manuscript submitted for publication: Rosa León Zayas, Mark Novotny, Sheila Podell, Charles M. Shepard, Eric Berkenpas, Sergey Nikolenko, Pavel Pevzner, Roger S. Lasken and Douglas H. Bartlett. 'Microbial Metabolic Properties at Greater Than 8,000 Meters Depth Within the Puerto Rico Trench Inferred From Single Cell Genomics' with permission from all coauthors

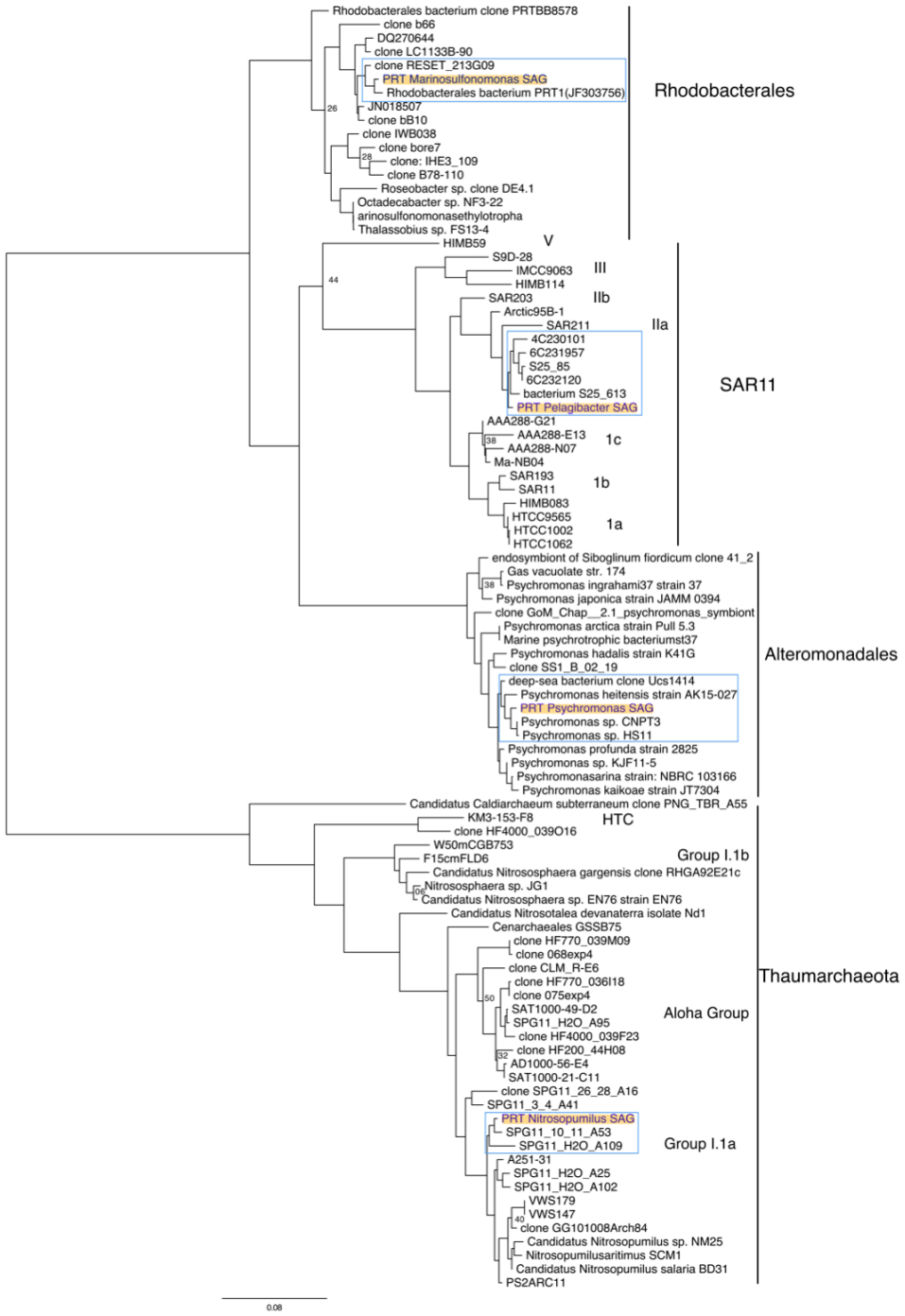

Figure 2.1 An unrooted maximum likelihood phylogenetic tree of the 16S rRNA gene from four single amplified genomes (SAGs) and related cultured and uncultured organisms is shown. The SAG names are highlighted in orange, and blue boxes denote deep-sea associated sequences. Phylogenetic divisions within groups are annotated. Scale bar represents 0.08 changes per position. The displayed confidence values are those that are 50% or lower.

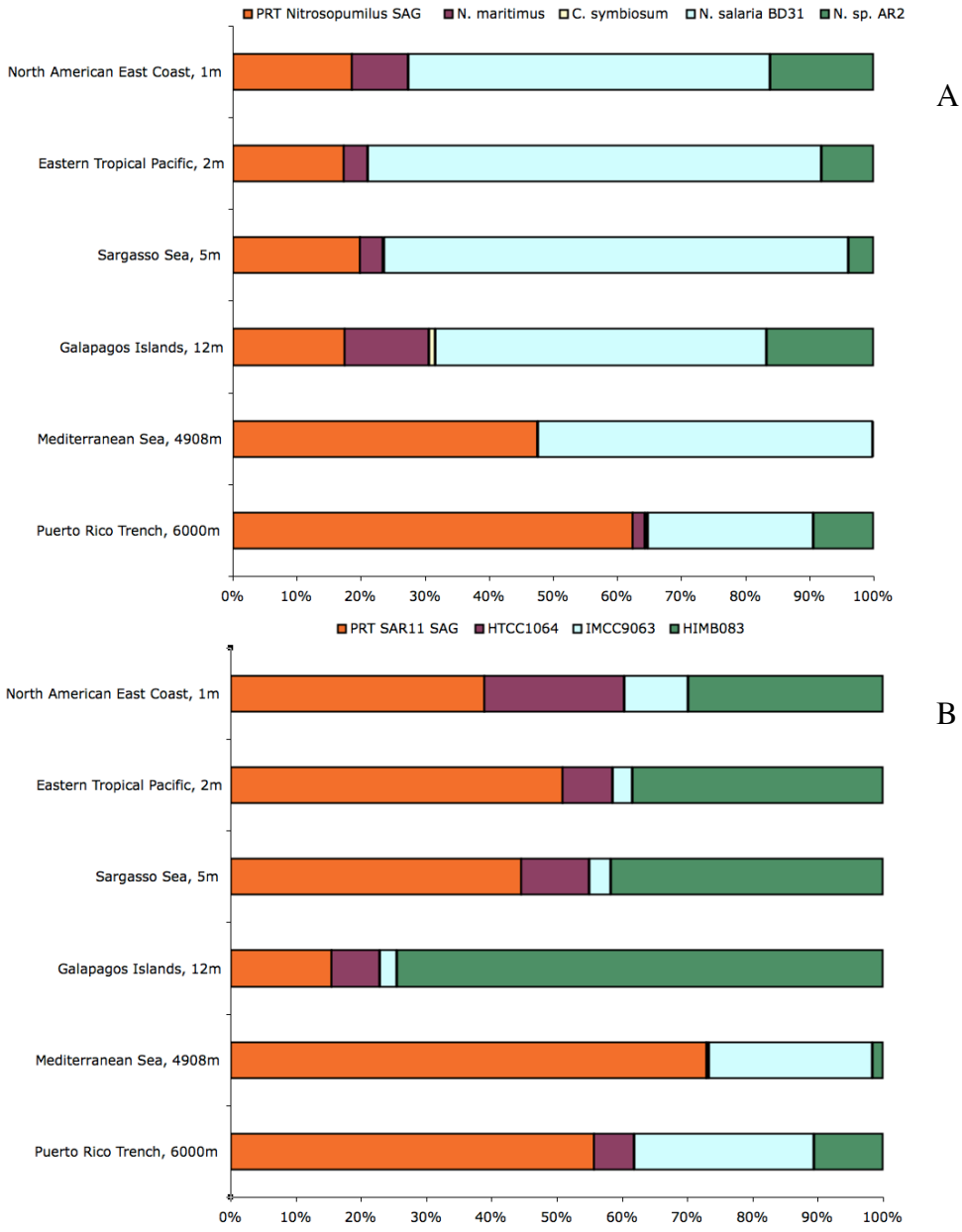

Figure 2.2 The relative abundance of the PRT *Nitrosopumilus* SAG and the PRT SAR11 SAG as assessed by reciprocal best blast (RBB) analysis shows a trend preferentially recruiting reads from metagenome data sets associated with deep-sea environments when compared to surface metagenomes. Metagenomes are displayed on the y-axis and the x-axis displays the percent of top hits during recruitment. A) PRT *Nitrosopumilus* SAG and related genomes: *N. maritimus - Nitrosopumilus maritimus SCM1*, *C. symbiosum - Cenarchaeum symbiosum*, *N. salaria BD31- Candidatus (Cand.) Nitrosopumilus salaria BD31 and N. sp AR2 – Cand. Nitrosopumilus sp. AR2*. B) PRT SAR11 SAG and related genomes: HTCC1062 – *Pelagibacter ubique* SAR11 HTCC1062, IMCC9063 - *Cand. Pelagibacter sp.* IMCC9063, HIMB083 – *Cand. Pelagibacter-like* (SAR11) HIMB083. RBB were normalized based on the genome size of the analyzed genome and total number of reads of the analyzed metagenome.

Table 2.1 Genomic characterization of 4 PRT SAGs.
Gene count, number of contigs, GC%, COG count and coding region % was obtained from the IMG web application. The estimated genome size was calculated based on the % completion and the size of the MCRG. The SAG genome size comparisons to their MCRGs are reported as – and the % difference, reflecting their smaller size respectively to the MCRG. The sign (+ or -) in the transposase column and coding region % indicate whether the SAG number is more (+) or less (-) than their respective MCRG.

Genome Name	Seq Genome	Est Genome Size	Gene Count	% Genome Completeness	GC %	16S rRNA	COG Count	Coding region %	Transposases
PRT *Marinosulfonomonas* SAG	1736454	2255135 (- , 34%)	1980	77	52%	1	1904	81.41 (-)	162 (+)
PRT *Nitrosopumilus* SAG	660313	917101 (- , 43%)	832	72	35%	1	682	80.59 (-)	1 (+)
PRT SAR11 SAG	700642	973113 (- , 25%)	704	72	30%	1	899	93.08 (-)	0
PRT *Psychromonas* SAG	920308	1559844 (- , 65%)	975	59	38%	3	1125	63.89 (-)	25 (-)

Table 2.2 Unique metabolic properties of the SAG genomes
Metabolic potential is listed as novel when it is not found in their most closely related genome (MCRG). The name of the predicted function, COG ID and BLASTP matches are presented.

Single Amplifed Genome	Function	Function ID	Top BLAST match
PRT *Nitrosopumilus*			
	FOG: PAS/PAC domain	COG2202	signal transduction histidine kinase, with PAS, phosphoacceptor and ATP binding domain [Candidatus Nitrososphaera gargensis]
	NhaD type sodium/proton-antiporters	COG1055	putative arsenical pump membrane protein [Nitrososphaera viennensis EN76]
	Lipoate-protein ligase	COG0095	lipoate-protein ligase A [Aciduliprofundum sp. MAR08-339]
	Lipoate synthase	COG0320	lipoyl synthase [Corallococcus coralloides]
	Glycine cleavage system H protein	COG0509	glycine cleavage system protein H [Dictyoglomus thermophilum]
	Glycine dehydrogenase subunit 1	COG0403	glycine dehydrogenase subunit 1 [Chloroherpeton thalassium]
	Glycine dehydrogenase subunit 2	COG1003	glycine dehydrogenase subunit 2 [Carboxydothermus hydrogenoformans]
	Aminomethyltransferase	COG0404	glycine cleavage system protein T [Kosmotoga olearia]
	Dihydrolipoamide dehydrogenase	COG1249	dihydrolipoamide dehydrogenase [Thermotoga lettingae]
	Urease subunit gamma	COG0831	urease subunit gamma [Nitrosopumilus sp. AR]
	Urease accessory protein	COG2371	Urease accessory protein [Cenarchaeum symbiosum]
	Urease accessory protein	COG0830	Urease accessory protein [Cenarchaeum symbiosum]
	Urease accessory protein	COG0378	urease accessory protein UreG [Cenarchaeum symbiosum]
	Urease accessory protein	COG0829	urease accessory protein UreD [Candidatus Nitrososphaera gargensis]
	Urease subunit alpha	COG0804	urease subunit alpha [Cenarchaeum symbiosum]
	Urease subunit beta	COG0832	urease subunit beta [Nitrososphaera viennensis EN76]
	Nitrogen regulatory protein P-II 2	K0:K04752	hypothetical protein [Candidatus Nitrosoarchaeum limnia]
	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acylt	COG1044	acetyltransferase [Candidatus Nitrosopumilus salaria]
	3-oxoacyl-[acyl-carrier-protein] synthase III	COG0332	3-oxoacyl-ACP synthase [Microcystis aeruginosa]
	Aquaporin-4	COG0580	glycerol transporter [Candidatus Nitrosoarchaeum limnia]
PRT *SAR11*			
	6-phosphofructokinase	COG0205	6-phosphofructokinase [alpha proteobacterium HIMB59]
	Pyruvate kinase	COG0469	pyruvate kinase [Hydrogenivirga sp. 128-5-R1-1]
	Taurine dioxygenase	COG2175	taurine catabolism dioxygenase TauD [Candidatus Pelagibacter sp. HTCC7211]
	TRAP-type C4-dicarboxylate transport system, large	COG1593	TRAP transporter, DctM subunit [[Clostridium bolteae]
	TRAP-type C4-dicarboxylate transport system, peripl	COG1638	DctP family TRAP transporter solute receptor [[Clostridium bolteae]
	TRAP-type C4-dicarboxylate transport system, small	COG3090	C4-dicarboxylate ABC transporter permease [Thermosinus carboxydivorans]
	TRAP-type uncharacterized transport system, peripla	COG2358	C4-dicarboxylate ABC transporter substrate-binding protein [alpha proteobacterium HIMB114]
	TRAP-type uncharacterized transport system, fused p	COG4666	C4-dicarboxylate ABC transporter [alpha proteobacterium HIMB114]
	ABC-type Mn/Zn transport systems, ATPase compon	COG1121	zinc transporter [alpha proteobacterium HIMB5]
	ABC-type sugar transport system, ATPase component	COG1129	sugar ABC transporter ATP-binding protein [Rhizobium sp. CF142]
	Ribose/xylose/arabinose/galactoside ABC-type transp	COG1172	Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components [Candidatus Pelagibacter ubique HIMB058]
	ABC-type sugar transport system, periplasmic compo	COG1879	LacI family transcriptional regulator [Kiloniella laminariae]
	ABC-type Zn2+ transport system, periplasmic compo	COG4531	zinc transporter [Candidatus Pelagibacter sp. HTCC7211]
	maltose/maltodextrin transport system ATP-binding p	COG3839	sugar ABC transporter ATP-binding protein [Rhizobium sp. CF142]
	Na+/melibiose symporter and related transporters	COG2211	symporter [alpha proteobacterium HIMB59]
	3-oxoacyl-[acyl-carrier-protein] synthase III	COG0332	hypothetical protein [Mariprofundus ferrooxydans]
PRT *Marinosulfonomonas*			
	Taurine dioxygenase	COG2175	gamma-butyrobetaine dioxygenase [Labrenzia sp. DG1229]
	Aquaporin Z	COG0580	aquaporin [Rhodobacter sphaeroides]
	Thiamine monophosphate synthase	COG0352	thiamine-phosphate pyrophosphorylase [Ahrensia sp. 13_GOM-1096m]
	Thiamine biosynthesis protein ThiC	COG0422	phosphomethylpyrimidine synthase [Pseudovibrio sp. FO-BEG1]
	Membrane-associated lipoprotein involved in thiamin	COG1477	thiamine biosynthesis protein ApbE [Pannonibacter phragmitetus]
	Hydroxymethylpyrimidine/phosphomethylpyrimidine	COG0351	hydroxymethylpyrimidine kinase [Sulfitobacter sp. MM-124]
	thiamine biosynthesis ThiG	KO:K03149	thiazole synthase [Sulfitobacter sp. MM-124]
	Biotin synthase and related enzymes	COG0502	biotin synthase [Ruegeria sp. R11]
	Dethiobiotin synthetase	COG0132	dithiobiotin synthetase [Phaeobacter gallaeciensis]
	Cobalamin biosynthesis protein CbiG	COG2073	precorrin-3B methylase [Roseobacter sp. AzwK-3b]
	Nitrous oxide reductase	COG4263	nitrous-oxide reductase [Ruegeria lacuscaerulensis]
	Regulator of nitric oxide reductase transcription	COG3901	FMN-binding protein [Roseobacter sp. SK209-2-6]
	Collagenase U32	COG0826	peptidase U32 [Phaeobacter arcticus]
	Co/Zn/Cd cations	COG0053	ABC transporter permease [Rhodobacter sp. SW2]
	Na+/H+ antiport	COG2111	cation:proton antiporter [Rhodobacteraceae bacterium HTCC2150]
	ABC-type transporters for multidrugs	COG0842	ABC-type multidrug transport system, permease component [Thalassobacter arenae]
	ABC-type transporters for dipeptide/oligopeptide/nic	COG1124	Oligopeptide transport ATP-binding protein OppF [Thalassobacter arenae]
	ABC-type transporters for long-chain fatty acids	COG1133	transporter [Candidatus Pelagibacter sp. HTCC7211]
	ABC-type transporters for 2-aminoethylphosphonate	COG1178	iron ABC transporter permease [Labrenzia sp. DG1229]
PRT *Psychromonas*			
	Ni,Fe-hydrogenase I small subunit	COG1740	quinone-reactive Ni/Fe-hydrogenase small chain [Shewanella halifaxensis]
	Ni,Fe-hydrogenase I large subunit	COG0374	hydrogenase 2 large subunit [Shewanella frigidimarina]
	Ni,Fe-hydrogenase I cytochrome b subunit	COG1969	hydrogenase [Shewanella loihica]
	Nitrate reductase cytochrome c-type subunit NapB	COG3043	nitrate reductase [Psychromonas sp. CNPT3]
	Periplasmic nitrate reductase chaperone NapD	COG3062	sorbose reductase [Psychromonas sp. CNPT3]
	Periplasmic nitrate reductase maturation protein NapI	COG1149	ferredoxin [Psychromonas sp. CNPT3]
	Trimethylamine N-oxide reductase system, NapE	COG4459	TorE protein [Psychromonas sp. CNPT3]
	ABC-type multidrug transport system, permease com	COG0842	ABC transporter [Psychromonas sp. CNPT3]
	Chemotaxis signal transduction protein	COG0835	nitrate/nitrite sensor protein NarQ [Psychromonas sp. CNPT3]

Table 2.3 Genes unique to the Puerto Rico Trench metagenome
SAG sequence reads that recruited to the Puerto Rico Trench metagenome but did not recruit to the other metagenome samples (GOS, NHRI/PCPA/FIGI/SARA; Venter *et al*, 2004; Rusch *et al*, 2007, HOT 4000 m; DeLong *et al*, 2006 and DeepMed 3000 m; Martin-Cuadrado *et al*, 2007 – see Supplementary figure 2.7) were identified using FR-HIT (Niu et al., 2012). The results are presented according to SAG genome, SAG gene product name, COG category, and the best BLASTN match. The PRT SAR11 SAG did not have any genes that uniquely recruited to the PRT metagenome.

Product name	COG category	COG ID	Blast result protein
PRT_Marinosulfonomonas_SAG			
arsenite oxidase, small subunit	[C] Energy production and conversion	COG0723	Arsenite oxidase, small subunit domain-containing protein [Pseudovibrio sp. FO-BEG1]
nitrous oxide reductase apoprotein	[C] Energy production and conversion	COG4263	nitrous-oxide reductase [Ruegeria lacuscaerulensis]
Intracellular septation protein A	[D] Cell cycle control, cell division, chromosome partitioning	COG2917	multidrug transporter [Roseobacter sp. SK209-2-6]
Tryptophan 2,3-dioxygenase (vermilion)	[E] Amino acid transport and metabolism	COG3483	tryptophan 2,3-dioxygenase [Leisingera methylohalidivorans DSM 14336]
4-aminobutyrate aminotransferase and related aminotransferases	[E] Amino acid transport and metabolism	COG0160	4-aminobutyrate aminotransferase [Rhizobium leguminosarum]
ABC-type sugar transport systems, ATPase components	[G] Carbohydrate transport and metabolism	COG3839	ABC transporter ATP-binding protein [Roseovarius sp. TM1035]
hypothetical protein	[K] Transcription	COG2002	transcriptional regulator [Candidatus Nitrosopumilus salaria]
Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	[K] Transcription	COG2197	transcriptional regulator [Octadecabacter arcticus 238]
3-methyladenine DNA glycosylase	[L] Replication, recombination and repair	COG2818	DNA-3-methyladenine glycosylase 1 [Roseobacter litoralis Och 149]
Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family	[O] Posttranslational modification, protein turnover, chaperones	COG1975	hypothetical protein thalar_02047 [Thalassobacter arenae DSM 19593]
Peroxiredoxin	[O] Posttranslational modification, protein turnover, chaperones	COG0678	peroxiredoxin [Roseobacter litoralis Och 149]
hypothetical protein	[R] General function prediction only	COG1058	molybdenum cofactor biosynthesis protein [Roseobacter sp. CCS2]
hypothetical protein	-	-	membrane protein [Nitrosopumilus sp. SJ]
Polysaccharide lyase family 8, N terminal alpha-helical domain.	-	-	silent information regulator protein Sir2 [Elizabethkingia meningoseptica]
hypothetical protein	-	-	sulfatase [Jannaschia sp. CCS1]
Hemerythrin HHE cation binding domain*	-	-	LysR family HTH-type transcriptional regulator [Phaeobacter gallaeciensis 2.10]
hypothetical protein	-	-	MORN repeat protein [endosymbiont of Ridgeia piscesae]
PRT_Psychromonas_SAG			
Ferredoxin	[C] Energy production and conversion	COG0633	ferredoxin [Psychromonas sp. CNPT3]
Phosphotransferase system fructose-specific component IIB	[G] Carbohydrate transport and metabolism	COG1445	PTS system, fructose-specific IIBC component [Psychromonas sp. CNPT3]
hypothetical protein	[N] Cell motility	COG1344	flagellin [Psychromonas sp. CNPT3]
PRT_Nitrosupumilus_SAG			
hypothetical protein	[C] Energy production and conversion	COG3794	blue (type1) copper domain-containing protein [Candidatus Nitrosopumilus sp. AR2]
Histones H3 and H4	[B] Chromatin structure and dynamics	COG2036	transcription factor CBF/NF-Y/histone domain-containing protein [Nitrosopumilus maritimus]
hypothetical protein	[K] Transcription	COG2002	AbrB family transcriptional regulator [Nitrosopumilus maritimus SCM1]
hypothetical protein	[K] Transcription		transcriptional regulator [Candidatus Nitrosopumilus salaria]

Supplementary Material

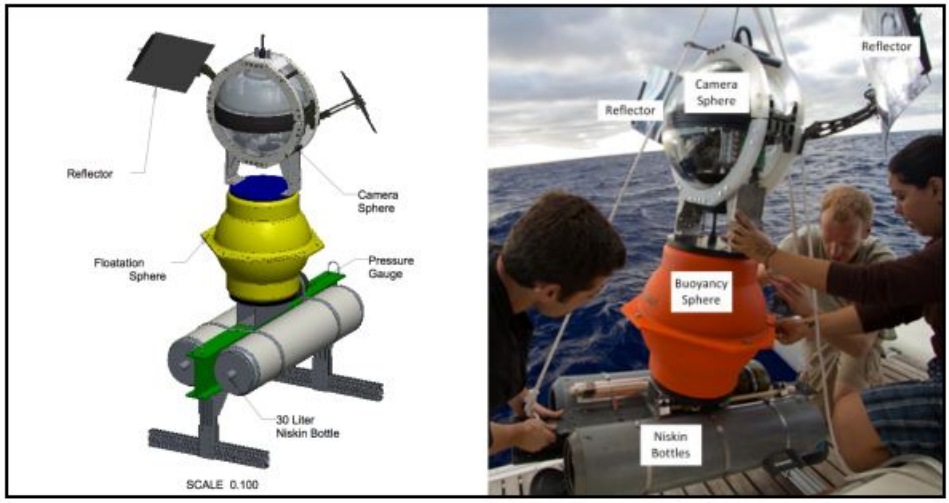

Figure S2.3. Free falling vehicle used to collect the ultra-deep seawater sample used in this study. The free falling vehicle (FFV) was designed and fabricated by National Geographic Remote Imaging. It contained a camera (HDR-XR520V, Sony, Tokyo, Japan), which captured high-definition video in 1080i format at 60 frames per second and recorded to an internal 240GB hard drive. Illumination was provided by two 3600 lumen LED arrays (BXRA-C4500, Bridgelux, Livermore, CA). A custom embedded computer commanded the camera and lighting based on pre-programmed timing. This system was encased in a polished 43 cm diameter, 2.5 cm thick borosilicate glass sphere with a depth rating of 12,000 m (Vitrovex, Nautilus Marine Service, GmbH, Buxtehude, Germany). A pair of external polycarbonate reflectors spread the illumination into the field of view of the camera. An external pressure gauge (DG25, Ashcroft, Stratford, CT) was used to measure the final depth that system achieved. To collect microbial samples the FV was fitted with a pair of baited 30 l Niskin water sampling bottles and an additional 43 cm sphere to provide additional buoyancy. The FFV was weighted with a 22 kg external steel ballast attached via a timed-release burnwire. Dissolving magnesium links provided a redundant release (A2, Neptune Marine Products, Port Townsend, WA). The FFV had an onboard radio beacon transmitter (MK8, Telonics, Mesa, AZ), which facilitated recovery using locating antennae. A backup ST-21H-200L Telonics satellite transmitter was used to determine position on the surface via the ARGOS satellite network. Video is available upon request.

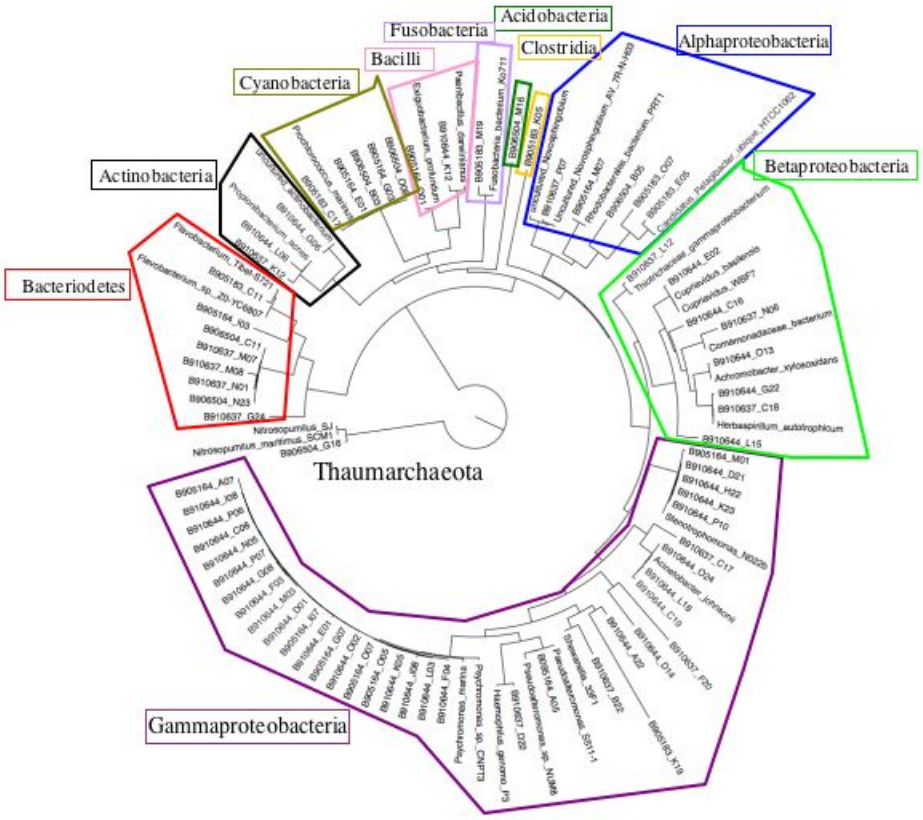

Figure S2.4 Phylogenetic tree of 16S rRNA gene sequences obtained from MDA amplifications. The maximum likelihood phylogenetic tree of 16S rRNA gene sequences obtained for 70 amplified SAG genomes is shown. Selected sequences are embedded within the amplified SAG for phylogenetic reference. The phylogenetic divisions of the relevant bacterial classes are highlighted.

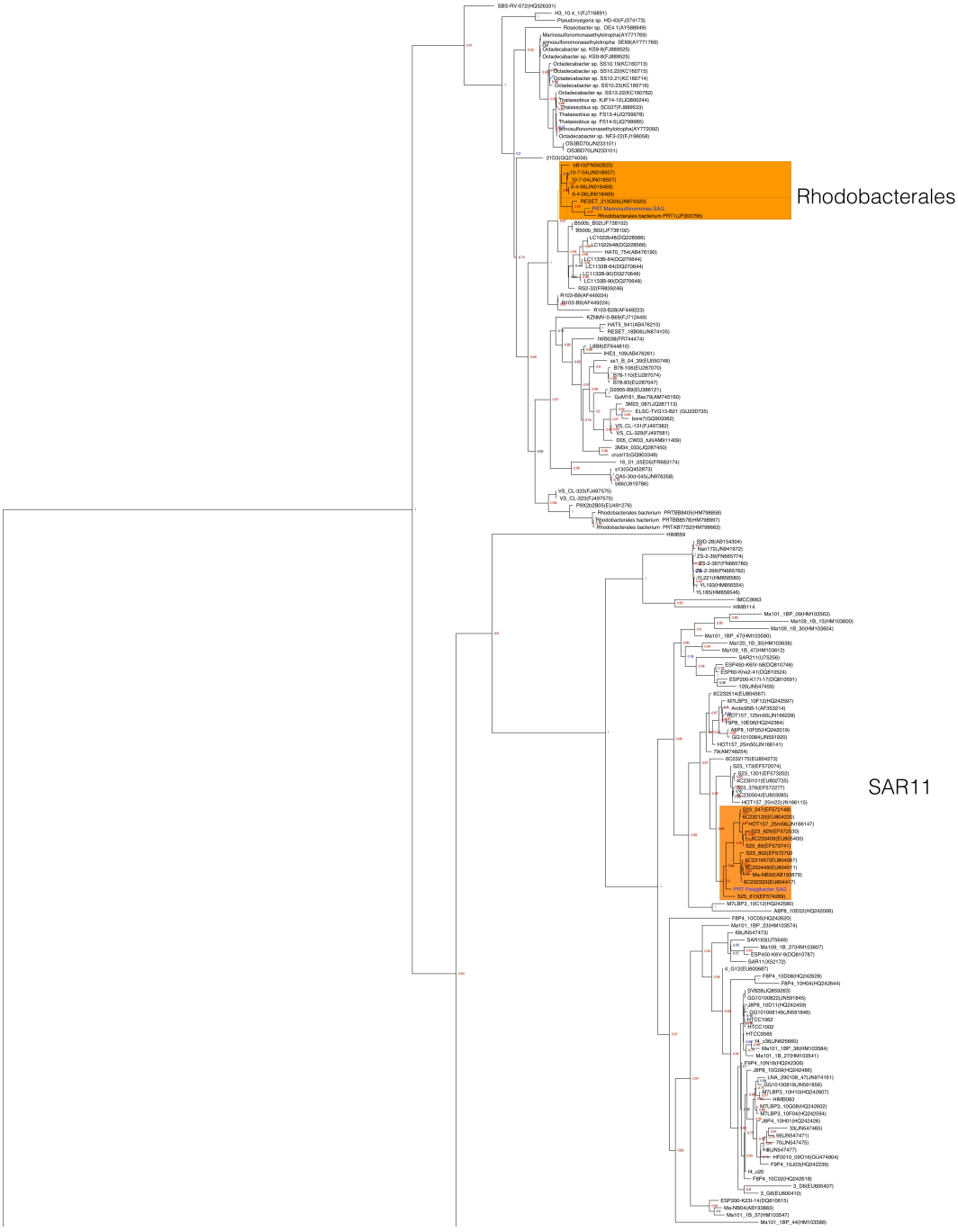

Figure S2.5 Expanded phylogenetic tree.
This figure displays an unrooted maximum likelihood phylogenetic tree of the 16S rRNA gene for the 4 SAGs and expanded selection of environmental and cultured organisms. Deep-sea associated clades are highlighted in orange. SAGs are colored in blue. Confidence values are displayed on the tree nodes and node colors identify numbers on a scale from 0 (blue) to 1 (red).

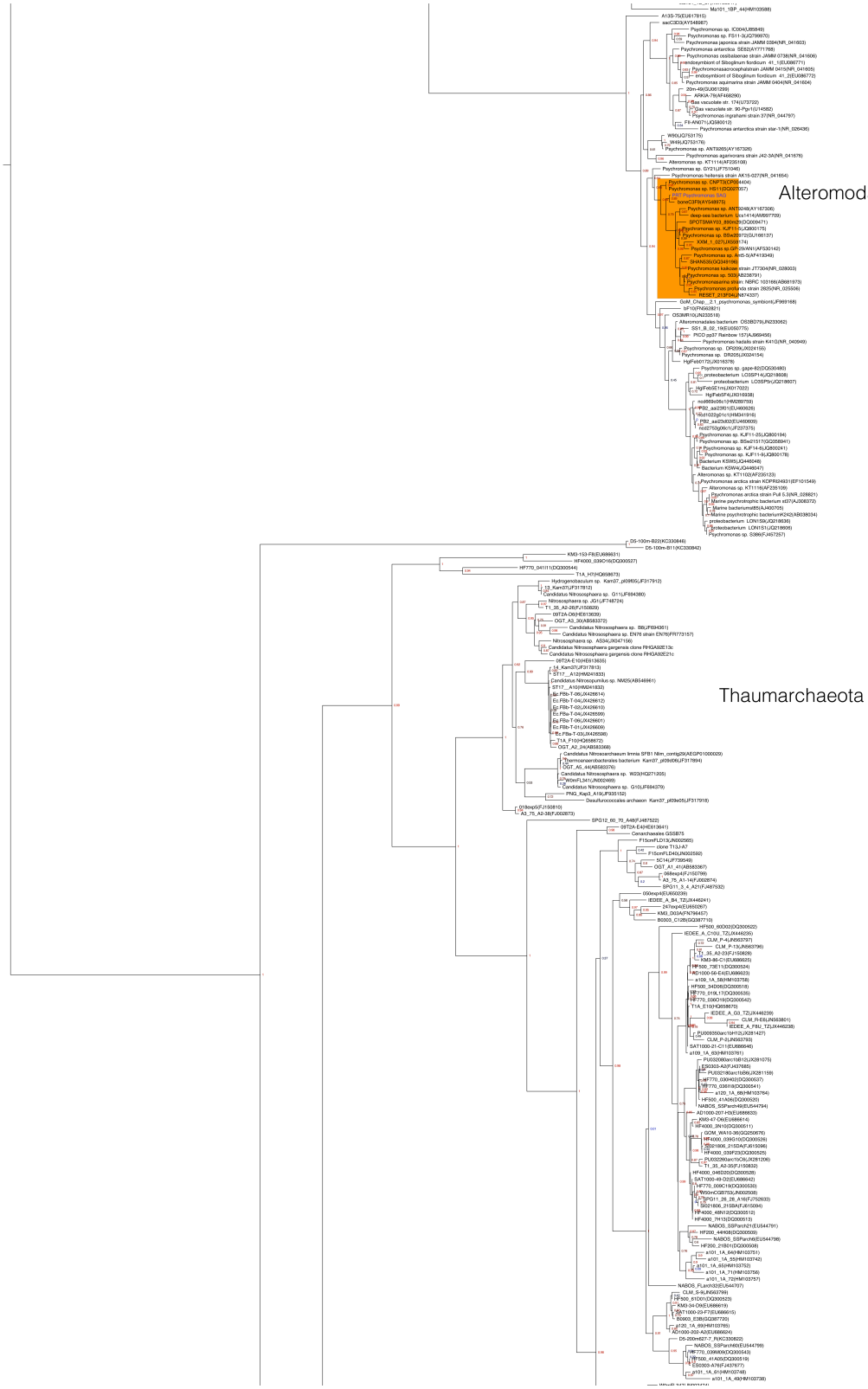

Figure S2.5 Expanded phylogenetic tree continued

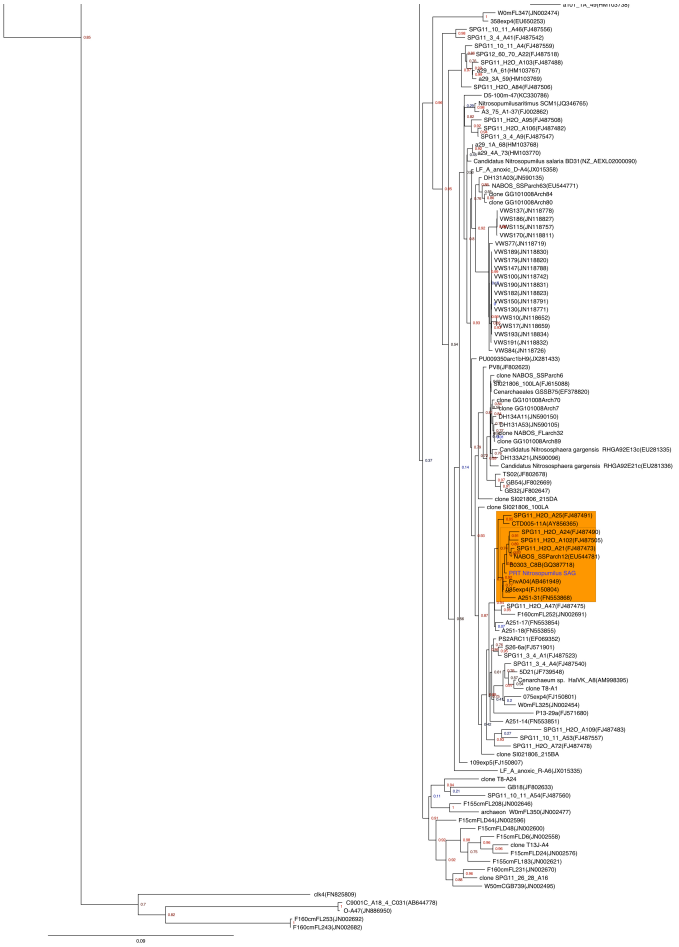

Figure S2.5 Expanded phylogenetic tree continued

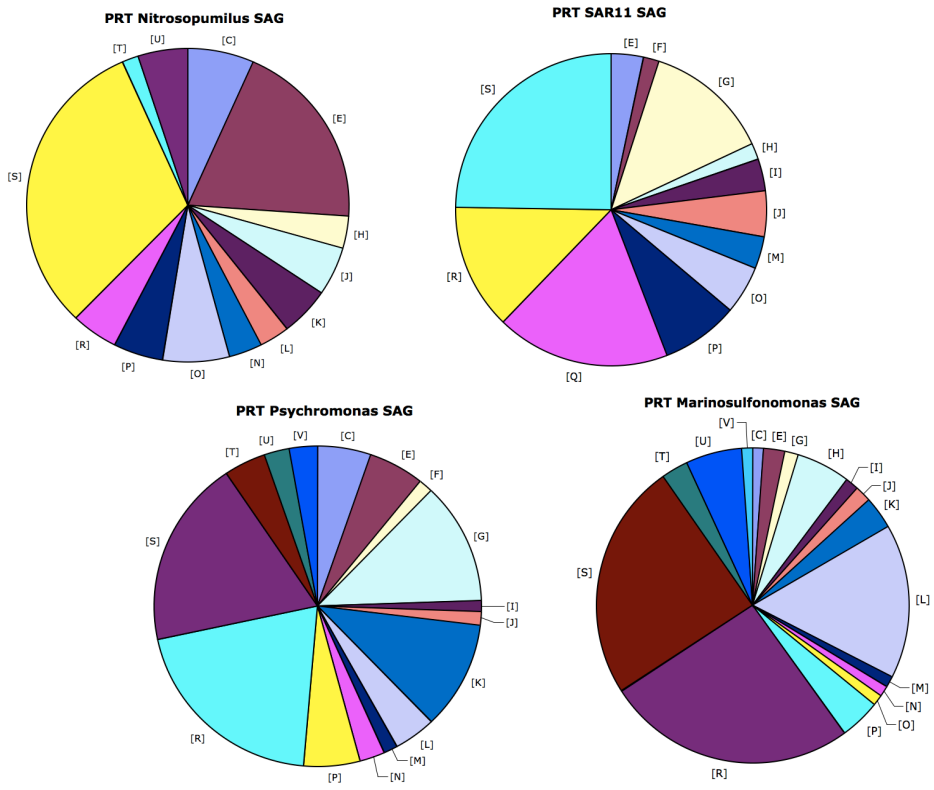

Figure S2.6 COG category distribution of unique genes when compared to SAGs most closely related genomes.
The COG categories and percentage for each SAG are displayed using pie charts. Category legend: [E] Amino acid transport and metabolism, [G] Carbohydrate transport and metabolism, [D] Cell cycle control, cell division, chromosome partitioning, [N] Cell motility, [M] Cell wall/membrane/envelope biogenesis, [B] Chromatin structure and dynamics, dynamics, [H] Coenzyme transport and metabolism, [V] Defense mechanisms, [C] Energy production and conversion, [S] Function unknown, [R] General function prediction only, [P] Inorganic ion transport and metabolism, [U] Intracellular trafficking, secretion, and vesicular transport, [I] Lipid transport and metabolism, [F] Nucleotide transport and metabolism, [O] Posttranslational modification, protein turnover, chaperones, [L] Replication, recombination and repair, [A] RNA processing and modification, [Q] Secondary metabolites biosynthesis, transport and catabolism, [T] Signal transduction mechanisms, [K] Transcription, [J] Translation, ribosomal structure and biogenesis.

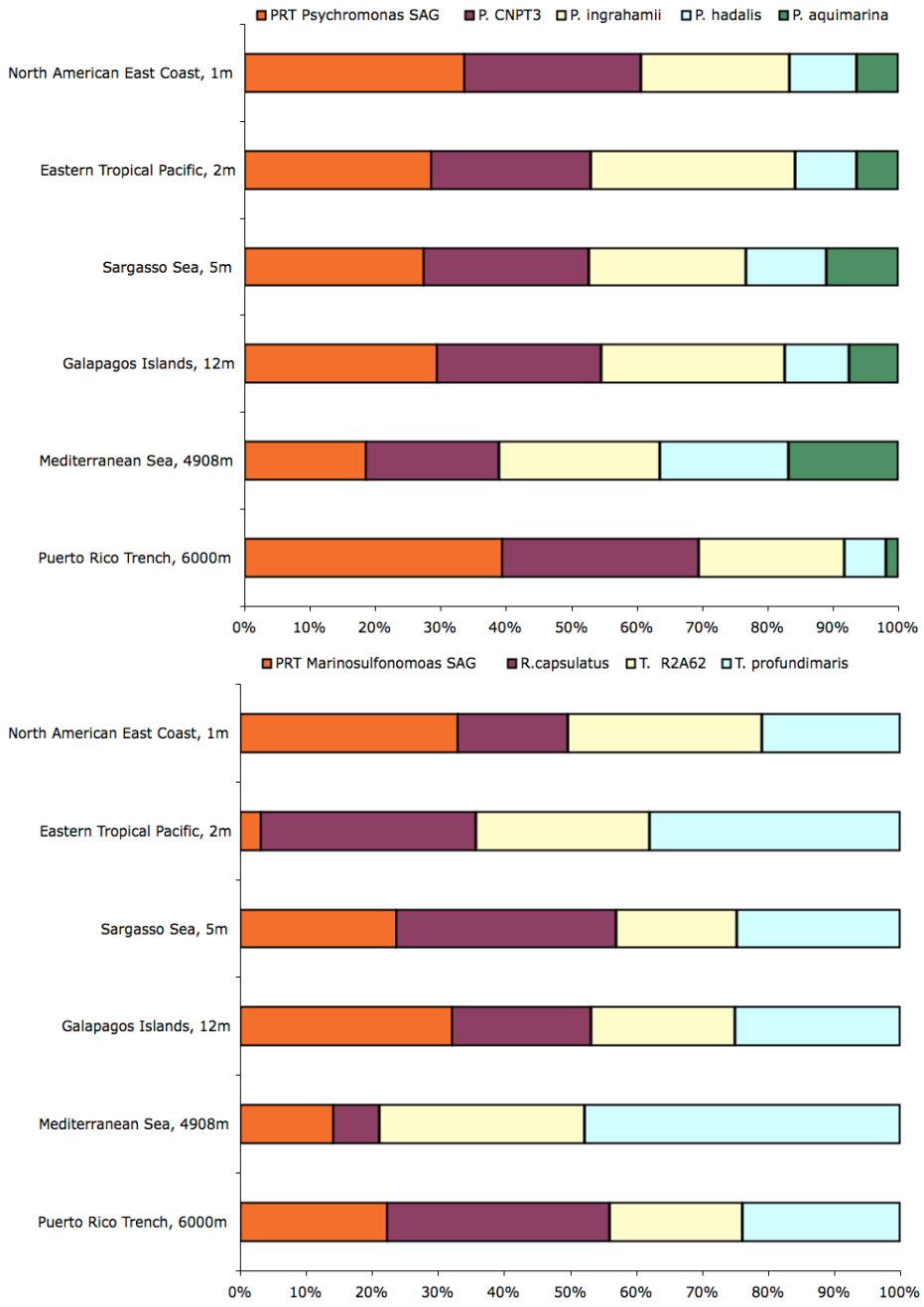

Figure S2.7 Reciprocal best blast for PRT Marinosulfonomonas SAG and PRT Psychromonas SAG

Relative abundance of PRT *Psychromonas* SAG and PRT *Marinosulfonomonas* SAG by reciprocal best blast analysis do not show a trend recruiting metagenome reads from deeper environments when compared to surface metagenomes. A) PRT Psychromonas SAG and related genomes: *P*. CNPT3 - *Psychromonas sp*. CNPT3, *P. hadalis* - *Psychromonas hadalis and P. aquamarina* - *Psychromonas aquamarina*. B) PRT *Marinosulfonomonas* SAG and related genomes: *R. capsulatus* –*Rhodobacter capsulatus*, *T*. R2A62 - *Thalassiobium sp*. *R2A62*, *T*. profundimaris –*Thalassiobium* profundimaris. RBB were normalized based on the genome size of the analyzed genome and total number of reads of the metagenome.

Table S2.4 List of metagenomes used in read recruitment analyses. The full list of metagenomes is indicated and highlighted are those use for comparisons shown in figure 2.2

Geographic Location	NICKNAME	Depth
Galapagos Islands	IIGI	0.10
Galapagos Islands	PCGI	0.20
Indian Ocean	IOEZ	0.30
North American East Coast	GUME	1.00
North American East Coast	BBME	1.00
North American East Coast	BBNS	1.00
North American East Coast	BFNS	1.00
North American East Coast	NOME	1.00
North American East Coast	NHRI	1.00
North American East Coast	BINY	1.00
North American East Coast	CMNJ	1.00
North American East Coast	DBNJ	1.00
North American East Coast	CHSC	1.00
Polynesia Archipelagos	RAFP	1.00
Eastern Tropical Pacific	DRCR	1.10
Polynesia Archipelagos	TLFP	1.20
Polynesia Archipelagos	CBFPa	1.40
Polynesia Archipelagos	CBFPb	1.40
Polynesia Archipelagos	MOFP	1.40
Indian Ocean	IOA	1.50
Indian Ocean	IOBa	1.50
Indian Ocean	IOBb	1.50
Indian Ocean	IOF	1.50
Indian Ocean	IOG	1.50
Indian Ocean	IOS	1.50
Indian Ocean	IOZ	1.50
Indian Ocean	IOWZ	1.50
Eastern Tropical Pacific	GUPA	1.60
Caribbean Sea	KWFL	1.70
Caribbean Sea	ROSA	1.70
Caribbean Sea	COPA	1.70
Galapagos Islands	WIGI	1.70
Tropical South Pacific	TSPE	1.70
Tropical South Pacific	FRPY	1.70
Eastern Tropical Pacific	EQPB	1.80
Tropical South Pacific	TSPA	1.80
Indian Ocean	CKILa	1.80
Indian Ocean	CKILb	1.80
Indian Ocean	IOC	1.80
Indian Ocean	IODa	1.80
Indian Ocean	IODb	1.80
Indian Ocean	IOE	1.80
Indian Ocean	IOSAa	1.80
Indian Ocean	IOSAb	1.80
Tropical South Pacific	TSPF	1.90
Tropical South Pacific	FRPX	1.90
Indian Ocean	IOYa	1.90
Indian Ocean	IOYb	1.90
North American East Coast	FXNS	2.00
Caribbean Sea	GMEX	2.00
Caribbean Sea	YUCA	2.00
Panama Canal	LGPA	2.00
Eastern Tropical Pacific	PCPA	2.00
Eastern Tropical Pacific	CICR	2.00
Galapagos Islands	NEGI	2.00
Galapagos Islands	FLGI	2.00
Tropical South Pacific	TSPB	2.00
Tropical South Pacific	TSPD	2.00
Tropical South Pacific	FRPZ	2.00
Indian Ocean	IORI	2.00
North American East Coast	CBVL	2.07
North American East Coast	NHNC	2.10
Galapagos Islands	SIGI	2.10
Galapagos Islands	SEGI	2.10
Galapagos Islands	CMGI	2.10
Galapagos Islands	DCGI	2.20
Tropical South Pacific	TSPC	2.20
Indian Ocean	IOX	2.20
Indian Ocean	IOM	2.80
Sargasso Sea	SARA	5.00
Sargasso Sea	SARA	5.00
Sargasso Sea	SARB	5.00
Sargasso Sea	SARB	5.00
Sargasso Sea	SARC	5.00
Sargasso Sea	SARD	5.00
Sargasso Sea	HYDA	5.00
Sargasso Sea	HYDB	5.00
Sargasso Sea	HYDC	5.00
Mediterranean Sea	ERR164407	5.00
Mediterranean Sea	ERR164409	5.00
Salish Sea	SRR944610	5.00
Salish Sea	SRR944614	5.00
Galapagos Islands	FIGI	12.00
North American East Coast	CBNJ	13.20
Galapagos Islands	RRGI	19.00
Tropical South Pacific	FRPW	30.00
Mediterranean Sea	SRR037008	50.00
Red Sea	SRR789380	50.00
Mediterranean Sea	ERR164408	56.00
Pacific Ocean, Chile Coast	SRR961671	70.00
Pacific Ocean, Chile Coast	SRR960580	70.00
Pacific Ocean, Chile Coast	SRR961675	110.00
Pacific Ocean, Chile Coast	SRR961673	110.00
Pacific Ocean, Chile Coast	SRR961677	200.00
Pacific Ocean, Chile Coast	SRR961676	200.00
Pacific Ocean, Chile Coast	SRR961679	1000.00
Pacific Ocean, Chile Coast	SRR961680	1000.00
Mid Atlantic Ridge	ERR133679	2320.00
Mid Atlantic Ridge	ERR133680	2320.00
Mid Atlantic Ridge	ERR133681	2320.00
Mediterranean Sea	DeepMed	3000.00
Hawaii Ocean Time Series	HOT	4000.00
Mediterranean Sea	SRR324677	4908.00
PRT		6000.00

Table S2.5 List of prophage predictions by ProphageFinder for the PRT *Marinosulfonomonas* SAG and the PRT *Psychromonas* SAG.
A) There are three different contigs unique to the PRT *Psychromonas* genome that were identified. Two of the loci are relatively small encoding five to six proteins predicted as phage-like, while the third loci has 21 identified phage-like proteins. The two small loci were homogeneous (Myoviridae and Inoviridae), while the third contig was chimeric. In the large contig 90% of the sequences were Myoviridae sequences, the other sequences belonged to the Podoviridae and Siphoviridae families. B) There were ten different contigs in the PRT *Marinosulfonomonas* genome that were identified as encoding phage-like proteins using ProphageFinder (Bose and Barber, 2006, http://bioinformatics.uwp.edu/~phage/ProphageFinder.php). Further analyses based on the Casjens et al. (2003) rational for recognizing prophage sequences in bacterial genomes identified seven loci as prophage-like loci. Out of 64 total predicted proteins within the prophage-like loci, Myoviridae was the most abundant prophage family followed by Inoviridae (78% and 16%, respectively). Most loci were chimeric in the sense that not all predicted proteins belonged to a phage from the same family.

PRT Psychromonas SAG

NODE_8018_length_6007

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_958083.1|Pag	3.00E-08	Enterobacteria phage PsP3	Myoviridae
ref|NP_536648.1|putative terminase, ATPase subunit	5.00E-06	Vibrio phage K139	Myoviridae
ref|NP_944196.1|hypothetical protein	4.00E-15	Aeromonas phage Aeh1	Myoviridae
ref|NP_892105.1|DNA adenine-methylase	5.00E-24	NP_892105.1	Myoviridae
ref|NP_536635.1|hypothetical protein	0.032	Vibrio phage K139	Myoviridae
ref|NP_536810.1|orf2(S)cox	0.003	Haemophilus phage HP2	Myoviridae

NODE_7634_length_23432

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_861801.1|Tk thymidine kinase	3.00E-50	Enterobacteria phage RB69	Myoviridae
ref|NP_061643.1|phi PVL ORF 20 and 21 homologue	0.2	Staphylococcus prophage phiPV83	Siphoviridae
ref|NP_758934.1|ORF43	3.00E-04	Vibrio phage VHML	Myoviridae
ref|NP_046780.1|gpE+E'	1.00E-13	Enterobacteria phage P2	Myoviridae
ref|NP_490623.1|hypothetical protein	5.00E-07	Pseudomonas phage phiCTX	Myoviridae
ref|NP_050643.1|major tail subunit	1.00E-45	Enterobacteria phage Mu	Myoviridae
ref|NP_599043.1|tail sheath protein	1.00E-09	Enterobacteria phage SfV	Myoviridae
ref|NP_050642.1|Hypothetical protein	1.00E-05	Enterobacteria phage Mu	Myoviridae
ref|NP_536654.1|putative tail completion protein	0.34	Vibrio phage K139	Myoviridae
ref|NP_043490.1|orf21	0.003	Haemophilus phage HP1	Myoviridae
ref|NP_046760.1|gpN	2.00E-46	Enterobacteria phage P2	Myoviridae
ref|NP_536822.1|scaffold	3.00E-23	Haemophilus phage HP2	Myoviridae
ref|NP_490600.1|predicted DNA-dependent ATPase terminase subunit	1.00E-105	Pseudomonas phage phiCTX	Myoviridae
ref|NP_958056.1|gp1	2.00E-57	Enterobacteria phage PsP3	Myoviridae
ref|NP_050974.1|P13	3.00E-20	Acyrthosiphon pisum bacteriophage APSE-1	Podoviridae
ref|NP_536668.1|putative tail fiber assembly protein	4.00E-18	Vibrio phage K139	Myoviridae
ref|NP_049863.1|gp37 long tail fiber, distal subunit	1.00E-25	Enterobacteria phage T4	Myoviridae
ref|NP_543104.1|hypothetical protein	8.00E-16	Enterobacteria phage phiP27	Myoviridae
ref|NP_050650.1|Hypothetical protein	1.00E-07	Enterobacteria phage Mu	Myoviridae
ref|NP_050649.1|putative baseplate assembly protein	5.00E-11	Enterobacteria phage Mu	Myoviridae
ref|NP_050648.1|putative tail protein	3.00E-05	Enterobacteria phage Mu	Myoviridae

NODE_1861_length_2942

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_047356.1|attachment protein	1.00E-09	Enterobacteria phage If1	Inoviridae
ref|NP_510890.1|structural protein	1.00E-33	Enterobacteria phage M13	Inoviridae
ref|NP_510889.1|structural protein	5.00E-11	Enterobacteria phage M14	Inoviridae
ref|NP_510888.1|phage assembly protein	8.00E-12	Enterobacteria phage M15	Inoviridae
ref|NP_510887.1|helix destabilising protein	1.00E-44	Enterobacteria phage M16	Inoviridae

Table S2.5 List of prophage predictions by ProphageFinder for the PRT *Marinosulfonomonas* SAG and the PRT *Psychromonas* SAG continued

PRT Marinosulfonomoas SAG

NODE_12835_length_19618

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_046906.1|gp11	0.012	Enterobacteria phage N15	Siphoviridae
ref|NP_758915.1|ORF22	2.00E-46	Vibrio phage VHML	Myoviridae
ref|NP_040583.1|capsid component	4.00E-25	Enterobacteria phage lambda	Siphoviridae
ref|NP_061501.1|ClpP protease	6.00E-20	Pseudomonas phage D3	Siphoviridae
ref|NP_758919.1|ORF26	4.00E-26	Vibrio phage VHML	Myoviridae
ref|NP_878213.1|gpV	3.00E-10	Enterobacteria phage WPhi	Myoviridae
ref|NP_891707.1|gp5.4 conserved hypothetical protein	9.00E-06	Enterobacteria phage RB49	Myoviridae
ref|NP_490616.1|predicted baseplate	2.00E-14	Pseudomonas phage phiCTX	Myoviridae
ref|NP_878215.1|gpJ	7.00E-29	Enterobacteria phage WPhi	Myoviridae
ref|NP_758926.1|ORF33	1.00E-13	Vibrio phage VHML	Myoviridae

NODE_12826_length_10964

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|YP_024909.1|putative endolysin	4.00E-14	Burkholderia phage BcepB1A	Myoviridae
ref|NP_758937.1|ORF46	4.00E-33	Vibrio phage VHML	Myoviridae
ref|NP_052256.1|Orf23; P2 X homolog; tail protein	7.00E-11	Enterobacteria phage 186	Myoviridae
ref|NP_046783.1|gpU	4.00E-11	Enterobacteria phage 186	Myoviridae
ref|NP_852558.1|hypothetical protein	0.098	Bacillus phage phBC6A52	
ref|NP_543099.1|putative tail protein	7.00E-64	Enterobacteria phage phiP27	Myoviridae
ref|NP_758932.1|ORF40	6.00E-10	Enterobacteria phage phiP27	Myoviridae
ref|NP_758931.1|ORF39	4.00E-71	Vibrio phage VHML	Myoviridae

NODE_12825_length_21568

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|YP_024909.1|putative	5.00E-13	Burkholderia phage BcepB1A	Myoviridae
ref|NP_700398.1|Probable tail fiber assembly protein	0.32	Salmonella phage ST64B	
ref|NP_932576.1|hinge connector of long tail fiber distal connector	3.00E-08	Aeromonas page 44RR2.8t	Myoviridae
ref|NP_536834.1|orf27	0.029	Haemophilus phage HP2	Myoviridae
ref|NP_061515.1|Orf19	4.00E-55	Pseudomonas phage D3	Siphoviridae
ref|NP_958590.1|putative major tail protein	0.002	Lactobacillus prophage Lj965	Siphoviridae
ref|NP_037704.1|Gp10	0.022	Enterobacteria phage HK97	Siphoviridae
ref|YP_164427.1|tail tape measure protein	0.14	Bacillus pahge BCJA1c	Siphoviridae
ref|NP_536362.1|putative major capcid protein	5.00E-52	Burkholderia phage phiE125	Siphoviridae
ref|NP_046900.1|gp5	2.00E-17	Enterobacteria phage N15	Siphoviridae
ref|YP_006584.1|putative portal protein	2.00E-23	Klebsiella pahge phiKO2	Siphoviridae
ref|NP_061498.1|terminase	1.00E-66	Pseudomonas phage D3	Siphoviridae
ref|NP_817455.1|gp5	9.00E-07	Mycobacterium phage Cjw1	Siphoviridae
ref|NP_690803.1|ORF19	1.00E-16	Bacillus phage phi105	Siphoviridae

NODE_12735_length_28507

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_795414.1|tail tapemeasure protein	0.31	Streptococcus pyogenes phage 315.1	
ref|NP_665964.1|putative plasmid partitioning protein Soj	0.006	Natrialba phage PhiCh1	Myoviridae
ref|NP_040579.1|G IV protein	2.00E-16	Enterobacteria phage Ike	Inovirus
ref|NP_944302.1|conserved phage protein	0.24	Burkholderia phage Bcep22	Podoviridae
ref|YP_024705.1|gp32	4.00E-35	Burkholderia phage BcepMu	Myoviridae
ref|YP_024707.1|gp34	1.00E-40	Burkholderia phage BcepMu	Myoviridae

NODE_12555_length_14011

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_899326.1|gp44	0.015	Vibrio phage KVP40	Myoviridae
ref|YP_164042.1|hypothetical protein	3.00E-29	Pseudomonas phage B3	Siphoviridae
ref|NP_852508.1|hypothetical protein	1.00E-10	Bacillus phage phBC6A51	
ref|NP_852509.1|hypothetical protein	4.00E-14	Bacillus phage phBC6A51	
ref|YP_024675.1|gp02	2.00E-11	Burkholderia phage BcepMu	Myoviridae
ref|NP_758912.1|ORF19	2.00E-09	Vibrio phage VHML	Myoviridae
ref|NP_542315.1|unknown	6.00E-17	Sinorhizobium phage PBC5	
ref|NP_046793.1|Orf82	3.00E-05	Enterobacteria phage P2	Myoviridae
ref|NP_050630.1|hypotheical protein	8.00E-04	Enterobacteria phage Mu	Myoviridae
ref|NP_050631.1|Hypothetical protein	1.00E-06	Enterobacteria phage Mu	Myoviridae
ref|YP_024701.1|gp28	1.00E-112	Burkholderia phage BcepMu	Myoviridae
ref|NP_938234.1|portal protein	1.00E-52	Pseudomonas phage D3112	Siphoviridae
ref|NP_050634.1|virion morphogenesis late F orf	7.00E-37	Enterobacteria phage Mu	Myoviridae

NODE_1255_length_17609

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|NP_695109.1|putative structural protein	4.00E-06	Streptococcus phage O1205	Siphoviridae
ref|NP_690787.1|immunity repressor	0.008	Bacillus phage phi105	Siphoviridae
ref|NP_076640.1|anti-repressor	1.00E-34	Lactococcus phage bIL286	Siphoviridae
ref|NP_958580.1|putative portal protein	0.005	Lactobacillus prophage Lj965	Siphoviridae
ref|YP_112491.1|terminase large subunit	3.00E-80	Flavobacterium phage 11b	Siphoviridae
ref|YP_112495.1|conserved hypothetical protein	4.00E-08	Flavobacterium phage 11b	Siphoviridae
ref|NP_047144.1|e29	0.031	Lactococcus phage bIL170	Siphoviridae
ref|NP_859332.1|hypothetical protein	0.15	Stx2 converting phage II	Siphoviridae
ref|NP_859216.1|hypothetical protein	8.00E-19	Stx1 converting phage	Siphoviridae
ref|NP_899374.1|conserved hypothetical protein	6.00E-04	Vibrio phage KVP40	Myoviridae

NODE_12447_length_14230

Best Blast Hit	Evalue	Taxanomic homolog	Family
ref|YP_164269.1|hypothetical protein	2.00E-32	Pseudomonas phage F116	Podoviridae
ref|NP_848233.1|endolysin	0.043	Enterobacteria phage epsilon15	Podoviridae
ref|NP_803710.1|ORF144	1.00E-04	Pseudomonas phage phiKZ	Myoviridae
ref|NP_438136.1|hypothetical protein	4.00E-21	Temperate phage phiNIH1.1	Siphoviridae
ref|NP_758915.1|ORF22	2.00E-61	Vibrio phage VHML	Myoviridae
ref|NP_758916.1|ORF23	6.00E-20	Vibrio phage VHML	Myoviridae
ref|NP_543091.1|putative prohead protease	0.13	Enterobacteria phage phiP27	Myoviridae
ref|YP_006586.1|major capsid head protein precursor	0.28	Klebsiella phage phiKO2	Siphoviridae
ref|NP_040589.1|head-tail joining	0.019	Enterobacteria phage lambda	Siphoviridae

Table S2.6 sRNAs annotated by the IMG platform for all four SAGs.

Gene ID	Locus Type	Gene Product Name	Coordinates	Length
PRT *Marinosulfonomonas* SAG				
2518659305	miscRNA	Long range pseudoknot	662..988(-)	327
2518660815	miscRNA	ALIL pseudoknot	338..454(+)	117
2518660819	miscRNA	ALIL pseudoknot	338..454(+)	117
2518661456	miscRNA	Alphaproteobacteria transfer-messenger RNA	3589..3940(+)	352
2518661915	miscRNA	Alpha operon ribosome binding site	3347..3446(-)	100
2518662159	miscRNA	C4 antisense RNA	20615..20700(-)	86
PRT *Nitrosopumilus* SAG				
2518654676	miscRNA	Selenocysteine transfer RNA	124..213(-)	90
PRT *Psychromonas* SAG				
2518653000	miscRNA	6S / SsrS RNA	2894..3077(+)	184
2518653107	miscRNA	Bacterial RNase P class A	6284..6644(+)	361
2518653278	miscRNA	Selenocysteine transfer RNA	439..528(+)	90
2518653710	miscRNA	Pseudomonas sRNA P26	15059..15116(-)	58
2518654150	miscRNA	Bacterial small signal recognition particle RNA	1104..1199(-)	96
PRT SAR11 SAG				
2518656674	miscRNA	Alphaproteobacteria transfer-messenger RNA	351..560(+)	210
2518656968	miscRNA	Bacterial small signal recognition particle RNA	1228..1326(-)	99
2518657042	miscRNA	SAR11_0636 sRNA	11190..11264(+)	75

Table S2.7 Unique genes and their respective COG, KEGG, EC and Pfam classifications

PRT SAR11										
DNA_length	**GC**	**Protein_length**	**Product_name**	**pfam**	**EC**	**KO_id**	**KO**	**COG_category**	**KEGG_module**	**COG_id**
276bp	0.17	91aa	Chorismate mutase	CM_2 [pfam01817]	-	-	-	[E] Amino acid transport and metabolism	-	COG1605
1038bp	0.26	345aa	Peptidylarginine deiminase and related enzymes	PAD_porph [pfam04371]	Agmatine deiminase.	K10536	agmatine deiminase [EC:3.5.3.12]0.00E+00	[E] Amino acid transport and metabolism	-	COG2957
876bp	0.33	291aa	methylthioadenosine phosphorylase (EC 2.4.2.28)	PNP_UDP_1 [pfam01048]	S-methyl-5'-thioadenosine phosphorylase.	K00772	5'-methylthioadenosine phosphorylase [EC:2.4.2.28]0.	[F] Nucleotide transport and metabolism	M00034: Methionine salvage pathway	COG0005
579bp	0.34	193aa	6-phosphofructokinase	PFK [pfam00365]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG0205
900bp	0.32	299aa	2-keto-myo-inositol dehydratase (EC 4.2.1.44)	AP_endonuc_2 [pfam01261]	Myo-inosose-2 dehydratase.	K03335	inosose dehydratase [EC:4.2.1.44]0.00E+00	[G] Carbohydrate transport and metabolism	-	COG1082
792bp	0.34	263aa	monosaccharide ABC transporter ATP-binding protein, (	ABC_tran [pfam00005]	Monosaccharide-transporting ATPase.	K02056	simple sugar transport system ATP-binding protein [EC	[G] Carbohydrate transport and metabolism	M00221: Putative simple sugar transport system	COG1129
1128bp	0.36	375aa	monosaccharide ABC transporter membrane protein, C	BPD_transp_2 [pfam02653]	-	K02057	simple sugar transport system permease protein0.00E-	[G] Carbohydrate transport and metabolism	M00221: Putative simple sugar transport system	COG1172
924bp	0.38	307aa	monosaccharide ABC transporter substrate-binding pro	Peripla_BP_4 [pfam13407]	-	K02058	simple sugar transport system substrate-binding protei	[G] Carbohydrate transport and metabolism	M00221: Putative simple sugar transport system	COG1879
654bp	0.26	217aa	Citrate lyase beta subunit	HpcH_HpaI [pfam03328]	Citryl-CoA lyase.	K01644	citrate lyase subunit beta / citryl-CoA lyase [EC:4.1.3.6	[G] Carbohydrate transport and metabolism	-	COG2301
792bp	0.34	263aa	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase	HpcH_HpaI [pfam03328]	Lyases. Carbon-carbon lyases. Aldehyde-lyases.	K02510	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase [EC:4.	[G] Carbohydrate transport and metabolism	-	COG3836
1119bp	0.24	372aa	Protein of unknown function (DUF563).	DUF563 [pfam04577]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG4421
672bp	0.32	223aa	Pyrroloquinoline quinone (Coenzyme PQQ) biosynthesis	TENA_THI-4 [pfam03070]	Pyrroloquinoline-quinone synthase.	K06137	pyrroloquinoline-quinone synthase [EC:1.3.3.11]0.00E	[H] Coenzyme transport and metabolism	-	COG5424
276bp	0.24	91aa	3-oxoacyl-[acyl-carrier-protein] synthase III	ACP_syn_III_C [pfam08541]	-	-	-	[I] Lipid transport and metabolism	-	COG0332
996bp	0.24	331aa	Predicted acyltransferases	Acyl_transf_3 [pfam01757]	-	-	-	[I] Lipid transport and metabolism	-	COG1835
483bp	0.29	161aa	Predicted translation initiation factor 2B subunit, eIF-2	IF-2B [pfam01008]	-	-	-	[J] Translation, ribosomal structure and biogenesis	-	COG0182
345bp	0.23	114aa	ribonuclease P protein component, eubacterial	Ribonuclease_P [pfam00825]	Ribonuclease P.	K03536	ribonuclease P protein component [EC:3.1.26.5]1.30E-	[J] Translation, ribosomal structure and biogenesis	-	COG0594
345bp	0.22	114aa	hypothetical protein	-	-	-	-	[J] Translation, ribosomal structure and biogenesis	-	COG2913
876bp	0.28	291aa	Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.2	NTP_transferase [pfam00483]	Glucose-1-phosphate thymidylyltransferase.	K00973	glucose-1-phosphate thymidylyltransferase [EC:2.7.7.2	[M] Cell wall/membrane/envelope biogenesis	-	COG1209
537bp	0.23	178aa	dTDP-4-dehydrorhamnose 3,5-epimerase and related e	dTDP_sugar_isom [pfam00908]	dTDP-4-dehydrorhamnose 3,5-epimerase.	K01790	dTDP-4-dehydrorhamnose 3,5-epimerase [EC:5.1.3.13	[M] Cell wall/membrane/envelope biogenesis	-	COG1898
1719bp	0.28	572aa	Highly conserved protein containing a thioredoxin dom	DUF255 [pfam03190]	-	-	-	[O] Posttranslational modification, protein turnover, ch	-	COG1331
948bp	0.28	315aa	ADP-ribosylglycohydrolase.	ADP_ribosyl_GH [pfam03747]	Hydrolases. Glycosylases.	K05521	ADP-ribosylglycohydrolase [EC:3.2.-.-]0.00E+00	[O] Posttranslational modification, protein turnover, ch	-	COG1397
723bp	0.32	240aa	Predicted proteasome-type protease	Proteasome [pfam00227]	-	K07395	putative proteasome-type protease0.00E+00	[O] Posttranslational modification, protein turnover, ch	-	COG3484
2172bp	0.37	723aa	catalase/peroxidase HPI	peroxidase [pfam00141]	Catalase peroxidase.	K03782	catalase-peroxidase [EC:1.11.1.21]0.00E+00	[P] Inorganic ion transport and metabolism	M00039: Lignin biosynthesis, cinnamate => lignir	COG0376
804bp	0.3	267aa	ABC-type Mn2+/Zn2+ transport systems, permease co	ABC-3 [pfam00950]	-	K09816	zinc transport system permease protein0.00E+00	[P] Inorganic ion transport and metabolism	M00242: Zinc transport system	COG1108
747bp	0.3	248aa	ABC-type Mn/Zn transport systems, ATPase componen	ABC_tran [pfam00005]	Hydrolases. Acting on acid anhydrides; catalyzing trans	K09817	zinc transport system ATP-binding protein [EC:3.6.3.-]	[P] Inorganic ion transport and metabolism	M00242: Zinc transport system	COG1121
576bp	0.32	191aa	ABC-type Zn2+ transport system, periplasmic compone	SBP_bac_9 [pfam01297]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG4531
501bp	0.33	166aa	Uncharacterized copper-binding protein	Cupredoxin_1 [pfam13473]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG4454
798bp	0.3	265aa	Probable taurine catabolism dioxygenase	TauD [pfam02668]	Taurine dioxygenase.	K03119	taurine dioxygenase [EC:1.14.11.17]0.00E+00	[Q] Secondary metabolites biosynthesis, transport and	-	COG2175
876bp	0.32	291aa	Predicted amidohydrolase	CN_hydrolase [pfam00795]	N-carbamoylputrescine amidase.	K12251	N-carbamoylputrescine amidase [EC:3.5.1.53]0.00E+0	[R] General function prediction only	-	COG0388
897bp	0.29	298aa	Sphingosine kinase and enzymes related to eukaryotic	DAGK_cat [pfam00781]	-	-	-	[R] General function prediction only	-	COG1597
999bp	0.37	332aa	TRAP transporter solute receptor, TAXI family	NMT1 [pfam09084]	-	-	-	[R] General function prediction only	-	COG2358
264bp	0.25	87aa	Predicted Fe-S protein	DUF1289 [pfam06945]	-	-	-	[R] General function prediction only	-	COG3313
1977bp	0.33	658aa	TRAP transporter, 4TM/12TM fusion protein	DctM [pfam06808]	-	-	-	[R] General function prediction only	-	COG4666
1101bp	0.3	366aa	L-alanine-DL-glutamate epimerase and related enzyme	MR_MLE_N [pfam02746]	-	-	-	[R] General function prediction only	-	COG4948
900bp	0.29	299aa	Predicted permease, DMT superfamily	EamA [pfam00892]	-	-	-	[R] General function prediction only	-	COG5006
417bp	0.31	138aa	hypothetical protein	-	-	-	-	[R] General function prediction only	-	COG4666
243bp	0.3	80aa	Predicted membrane protein (DUF2061).	DUF2061 [pfam09834]	-	-	-	[S] Function unknown	-	COG3205
378bp	0.33	125aa	Uncharacterized conserved protein	Cyclophil_like [pfam04126]	-	K09143	hypothetical protein1.30E-20	[S] Function unknown	-	COG2164
576bp	0.22	191aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG4649
633bp	0.26	210aa	Uncharacterized conserved protein	DUF159 [pfam02586]	-	-	-	[S] Function unknown	-	COG2135
1443bp	0.31	480aa	Uncharacterized conserved protein	CP_ATPgrasp_1 [pfam04174]	-	-	-	[S] Function unknown	-	COG2308
855bp	0.3	284aa	Uncharacterized protein conserved in bacteria	Metallophos_2 [pfam12850]	-	-	-	[S] Function unknown	-	COG2908

Table S2.7 Unique genes and their respective COG, KEGG, EC and Pfam classifications continued

PRT Nitrosopumilus

DNA_length	GC	Protein_length	Product_name	pfam	EC	KO_id	KO	COG_category	KEGG_module	COG_category
1308bp	0.37	435aa	dihydrolipoamide dehydrogenase	Pyr_redox_dim [pfam0	Dihydrolipoyl dehydrogenase.	K00382	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	[C] Energy production and conversion	M00009: Citrate cycle (TCA cycle, Krebs cycle)	COG1249
1110bp	0.34	369aa	Redoxin.	Redoxin [pfam08534]	-	-	-	[C] Energy production and conversion	-	COG0526
432bp	0.37	143aa	succinate dehydrogenase subunit C (EC 1.3.5.1)	Sdh_cyt [pfam01127]	-	K00241	succinate dehydrogenase cytochrome b556 subunit	[C] Energy production and conversion	M00149: Succinate dehydrogenase, prokaryotes	COG2009
345bp	0.36	114aa	Succinate dehydrogenase, hydrophobic anchor subunit	Sdh_cyt [pfam01127]	-	K00241	succinate dehydrogenase cytochrome b556 subunit1.8(	[C] Energy production and conversion	M00009: Citrate cycle (TCA cycle, Krebs cycle)	COG2142
477bp	0.36	158aa	Bacterial transferase hexapeptide (six repeats).	Hexapep [pfam00132]	-	-	-	[E] Amino acid transport and metabolism	-	COG2171
381bp	0.31	126aa	glycine cleavage system H protein	GCV_H [pfam01597]	-	K02437	glycine cleavage system H protein1.20E-33	[E] Amino acid transport and metabolism	M00532: Photorespiration	COG0509
1254bp	0.36	417aa	Glycine cleavage system protein P (pyridoxal-binding),	GDC-P [pfam02347]	Glycine dehydrogenase (decarbo	K00282	glycine dehydrogenase subunit 1 [EC:1.4.4.2]0.00E+0(	[E] Amino acid transport and metabolism	-	COG0403
1056bp	0.35	351aa	glycine cleavage system T protein	GCV_T_C [pfam08669]	Aminomethyltransferase.	K00605	aminomethyltransferase [EC:2.1.2.10]0.00E+00	[E] Amino acid transport and metabolism	-	COG0404
1413bp	0.36	470aa	glycine dehydrogenase (decarboxylating) beta subunit	Aminotran_5 [pfam002	Glycine dehydrogenase (decarbo	K00283	glycine dehydrogenase subunit 2 [EC:1.4.4.2]0.00E+0(	[E] Amino acid transport and metabolism	-	COG1003
861bp	0.31	286aa	L-proline dehydrogenase (EC 1.5.99.8)	Pro_dh [pfam01619]	Proline dehydrogenase.	K00318	proline dehydrogenase [EC:1.5.99.8]0.00E+00	[E] Amino acid transport and metabolism	-	COG0506
861bp	0.31	286aa	L-proline dehydrogenase (EC 1.5.99.8)	Pro_dh [pfam01619]	Proline dehydrogenase.	K00318	proline dehydrogenase [EC:1.5.99.8]0.00E+00	[E] Amino acid transport and metabolism	-	COG0506
1983bp	0.41	660aa	transporter, SSS family	SSF [pfam00474]	-	-	-	[E] Amino acid transport and metabolism	-	COG0591
327bp	0.27	109aa	Urea amidohydrolase (urease) gamma subunit	Urease_gamma [pfam(	-	-	-	[E] Amino acid transport and metabolism	-	COG0831
1713bp	0.38	570aa	urease, alpha subunit	Amidohydro_1 [pfam0:	Urease.	K01428	urease subunit alpha [EC:3.5.1.5]0.00E+00	[E] Amino acid transport and metabolism	-	COG0804
369bp	0.36	122aa	urease, beta subunit	Urease_beta [pfam006	Urease.	K01429	urease subunit beta [EC:3.5.1.5]2.00E-41	[E] Amino acid transport and metabolism	-	COG0832
336bp	0.36	111aa	urease, gamma subunit	Urease_gamma [pfam(	Urease.	K01430	urease subunit gamma [EC:3.5.1.5]3.30E-40	[E] Amino acid transport and metabolism	-	COG0831
894bp	0.36	297aa	lipoate synthase	Radical_SAM [pfam040	Lipoyl synthase.	K03644	lipoic acid synthetase [EC:2.8.1.8]0.00E+00	[H] Coenzyme transport and metabolism	-	COG0320
753bp	0.33	250aa	Lipoate-protein ligase A	BPL_LplA_LipB [pfam0:	Lipoate--protein ligase.	K03800	lipoate-protein ligase A [EC:2.7.7.63]0.00E+00	[H] Coenzyme transport and metabolism	-	COG0095
168bp	0.36	55aa	LSU ribosomal protein L40E	-	-	K02927	large subunit ribosomal protein L40e1.10E-23	[J] Translation, ribosomal structure and biogenesis	M00177: Ribosome, eukaryotes	COG1552
270bp	0.38	89aa	RNA-binding protein involved in rRNA processing	Gar1 [pfam04410]	-	K07569	RNA-binding protein3.40E-35	[J] Translation, ribosomal structure and biogenesis	-	COG3277
303bp	0.37	100aa	RNase P subunit RPR2	Rpr2 [pfam04032]	Ribonuclease P.	K03540	ribonuclease P protein subunit RPR2 [EC:3.1.26.5]8.20	[J] Translation, ribosomal structure and biogenesis	-	COG2023
657bp	0.35	218aa	urease accessory protein UreG	cobW [pfam02492]	-	K03189	urease accessory protein0.00E+00	[K] Transcription	-	COG0378
321bp	0.41	107aa	Transposase and inactivated derivatives	DDE_3 [pfam13358]	-	-	-	[L] Replication, recombination and repair	-	COG3335
600bp	0.27	199aa	Tetratricopeptide repeat./TPR repeat.	TPR_2 [pfam07719]	-	-	-	[N] Cell motility	-	COG3063
369bp	0.33	123aa	TPR repeat.	TPR_11 [pfam13414]	-	-	-	[N] Cell motility	-	COG3063
273bp	0.34	90aa	Fe-S cluster protector protein	Iron_traffic [pfam0436	-	-	-	[O] Posttranslational modification, protein turnover, ch	-	COG2924
486bp	0.29	161aa	Urease accessory protein UreE	UreE_N [pfam02814]	-	K03187	urease accessory protein1.90E-36	[O] Posttranslational modification, protein turnover, ch	-	COG2371
702bp	0.31	233aa	Urease accessory protein UreF	UreF [pfam01730]	-	K03188	urease accessory protein0.00E+00	[O] Posttranslational modification, protein turnover, ch	-	COG0830
927bp	0.33	308aa	Urease accessory protein UreH	UreD [pfam01774]	-	K03190	urease accessory protein0.00E+00	[O] Posttranslational modification, protein turnover, ch	-	COG0829
372bp	0.33	124aa	ABC-type hemin transport system, periplasmic compon	Peripla_BP_2 [pfam01(	-	-	-	[P] Inorganic ion transport and metabolism	-	COG4558
1401bp	0.36	466aa	Na+/H+ antiporter NhaD and related arsenite permeas	CitMHS [pfam03600]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG1055
987bp	0.37	328aa	Phosphate/sulphate permeases	PHO4 [pfam01384]	-	K03306	inorganic phosphate transporter, PiT family0.00E+00	[P] Inorganic ion transport and metabolism	-	COG0306
471bp	0.35	156aa	Predicted Fe-S-cluster oxidoreductase	FliB [pfam03692]	-	-	-	[R] General function prediction only	-	COG0727
492bp	0.35	163aa	Predicted kinase	AAA_33 [pfam13671]	-	-	-	[R] General function prediction only	-	COG4639
741bp	0.31	246aa	condensin subunit ScpA	ScpA_ScpB [pfam0261	-	K05896	segregation and condensation protein A0.00E+00	[S] Function unknown	-	COG1354
195bp	0.35	64aa	conserved hypothetical protein	DUF2196 [pfam09962]	-	-	-	[S] Function unknown	-	COG4895
531bp	0.31	176aa	GINS complex protein.	Sld5 [pfam05916]	-	-	-	[S] Function unknown	-	COG1711
528bp	0.33	175aa	PAS domain S-box	PAS_9 [pfam13426]	-	-	-	[T] Signal transduction mechanisms	-	COG2202
171bp	0.37	56aa	Preprotein translocase subunit Sec61beta	Sec61_beta [pfam0391	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG4023
180bp	0.36	59aa	protein translocase SEC61 complex gamma subunit, ar	-	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG2443
339bp	0.35	113aa	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltran	Hexapep [pfam00132]	-	KOG14(	Translation initiation factor 2B, epsilon subunit (eIF-2B	[M] Cell wall/membrane/envelope biogenesis	-	COG1044
750bp	0.31	249aa	TPR repeat.	TPR_11 [pfam13414]	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG3063
1098bp	0.33	365aa	3-oxoacyl-[acyl-carrier-protein] synthase III	-	-	-	-	[I] Lipid transport and metabolism	-	COG0332
237bp	0.38	79aa	hypothetical protein	-	-	-	-	[K] Transcription	-	COG2002
279bp	0.37	93aa	hypothetical protein	-	-	-	-	[K] Transcription	-	COG2002
231bp	0.36	76aa	hypothetical protein	-	-	-	-	[L] Replication, recombination and repair	-	COG1107
405bp	0.33	134aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG1917
786bp	0.28	261aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG4372
1047bp	0.33	348aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG4955
1149bp	0.32	382aa	Uncharacterized NAD(FAD)-dependent dehydrogenases	Pyr_redox_2 [pfam079	-	-	-	[R] General function prediction only	-	COG0446
438bp	0.34	145aa	Uncharacterized conserved protein	DNase-RNase [pfam02	-	-	-	[S] Function unknown	-	COG1259
471bp	0.27	156aa	Uncharacterized conserved protein	DUF82 [pfam01927]	-	K09122	hypothetical protein2.30E-26	[S] Function unknown	-	COG1656
213bp	0.26	70aa	Uncharacterized conserved protein	DUF167 [pfam02594]	-	K09131	hypothetical protein2.90E-07	[S] Function unknown	-	COG1872
285bp	0.31	94aa	Uncharacterized conserved protein	MtN3_slv [pfam03083]	-	K15383	MtN3 and saliva related transmembrane protein4.50E-:	[S] Function unknown	-	COG4095
354bp	0.29	118aa	Uncharacterized conserved protein, contains double-str	Cupin_2 [pfam07883]	-	-	-	[S] Function unknown	-	COG1917
558bp	0.3	185aa	Uncharacterized protein conserved in archaea	-	-	K09723	hypothetical protein0.00E+00	[S] Function unknown	-	COG1711
753bp	0.35	250aa	Uncharacterized protein conserved in bacteria	DUF726 [pfam05277]	-	-	-	[S] Function unknown	-	COG4782

Table S2.7 Unique genes and their respective COG, KEGG, EC and Pfam classifications continued

PRT Psychromonas DNA_length	GC	Protein_length	Product_name	pfam	EC	KO_id	KO	COG_category	KEGG_module	COG_id
477bp	0.37	158aa	periplasmic nitrate reductase maturation protein NapF	Fer4_7 [pfam12838]	-	K02572	ferredoxin-type protein NapF0.00E+00	[C] Energy production and conversion	-	COG1149
1701bp	0.42	566aa	Ni,Fe-hydrogenase I large subunit	NiFeSe_Hases [pfam00374]	Hydrogen:quinone oxidoreductase.	K05922	quinone-reactive Ni/Fe-hydrogenase large subunit [EC:	[C] Energy production and conversion	-	COG0374
840bp	0.41	279aa	Ni/Fe-hydrogenase, b-type cytochrome subunit	Cytochrom_B_N [pfam00033]	-	K03620	Ni/Fe-hydrogenase 1 B-type cytochrome subunit0.00E-	[C] Energy production and conversion	-	COG1969
171bp	0.33	56aa	trimethylamine N-oxide reductase system, TorE protein	NapE [pfam06796]	-	K02571	periplasmic nitrate reductase NapE2.90E-15	[C] Energy production and conversion	-	COG4459
1233bp	0.38	410aa	sodium--glutamate symport carrier (gltS)	Glt_symporter [pfam03616]	-	K03312	glutamate:Na+ symporter, ESS family0.00E+00	[E] Amino acid transport and metabolism	-	COG0786
795bp	0.44	264aa	Asparaginase	Asparaginase_2 [pfam01112]	-	-	-	[E] Amino acid transport and metabolism	-	COG1446
1374bp	0.4	457aa	L-seryl-tRNA(Sec) selenium transferase (EC 2.9.1.1)	SelA [pfam03841]	L-seryl-tRNA(Sec) selenium transferase.	K01042	L-seryl-tRNA(Ser) seleniumtransferase [EC:2.9.1.1]0.0	[E] Amino acid transport and metabolism	-	COG1921
1473bp	0.38	490aa	amino acid/peptide transporter (Peptide:H+ symporter	PTR2 [pfam00854]	-	K03305	proton-dependent oligopeptide transporter, POT family	[E] Amino acid transport and metabolism	-	COG3104
528bp	0.32	175aa	hypoxanthine phosphoribosyltransferase (EC 2.4.2.8)	Pribosyltran [pfam00156]	Hypoxanthine phosphoribosyltransferase.	K00760	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]0	[F] Nucleotide transport and metabolism	-	COG0634
498bp	0.34	165aa	Phosphotransferase system cellobiose-specific compone	PTS_EIIC [pfam02378]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG1455
1401bp	0.39	466aa	Major Facilitator Superfamily.	MFS_1 [pfam07690]	-	K08218	MFS transporter, PAT family, beta-lactamase induction	[G] Carbohydrate transport and metabolism	-	COG0738
693bp	0.41	230aa	L-ribulose 5-phosphate 4-epimerase (EC 5.1.3.4)	Aldolase_II [pfam00596]	L-ribulose-5-phosphate 4-epimerase.	K03077	L-ribulose-5-phosphate 4-epimerase [EC:5.1.3.4]0.00E	[G] Carbohydrate transport and metabolism	-	COG0235
309bp	0.39	102aa	Phosphotransferase system cellobiose-specific compone	PTS_IIB [pfam02302]	Protein-N(pi)-phosphohistidine--sugar phosphotransfer	K02760	PTS system, cellobiose-specific IIB component [EC:2.7	[G] Carbohydrate transport and metabolism	M00275: PTS system, cellobiose-specific II component	COG1440
1197bp	0.37	398aa	mannose-6-phosphate isomerase, type 1 (EC 5.3.1.8)	PMI_typeI [pfam01238]	Mannose-6-phosphate isomerase.	K01809	mannose-6-phosphate isomerase [EC:5.3.1.8]0.00E+0	[G] Carbohydrate transport and metabolism	M00114: Ascorbate biosynthesis, plants, glucose-6P =>	COG1482
1317bp	0.4	438aa	Sugar phosphate permease	MFS_1 [pfam07690]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG2271
468bp	0.33	155aa	uncharacterized protein, YhcH/YjgK/YiaL family	DUF386 [pfam04074]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG2731
915bp	0.37	304aa	Glycosyl hydrolase family 92.	Glyco_hydro_92 [pfam07971]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG3537
285bp	0.32	95aa	Fucose dissimilation pathway protein FucU	RbsD_FucU [pfam05025]	-	-	-	[G] Carbohydrate transport and metabolism	-	COG4154
522bp	0.36	173aa	phosphatidylglycerophosphatase (EC 3.1.3.27)	PgpA [pfam04608]	Phosphatidylglycerophosphatase.	K01095	phosphatidylglycerophosphatase A [EC:3.1.3.27]0.00E	[I] Lipid transport and metabolism	-	COG1267
1866bp	0.39	621aa	selenocysteine-specific translation elongation factor Se	GTP_EFTU [pfam00009]	-	K03833	selenocysteine-specific elongation factor0.00E+00	[J] Translation, ribosomal structure and biogenesis	-	COG3276
279bp	0.36	92aa	Predicted transcriptional regulator with C-terminal CBS	HTH_3 [pfam01381]	-	-	-	[K] Transcription	-	COG3620
795bp	0.37	264aa	transcriptional regulator, IclR family	HTH_IclR [pfam09339]	-	-	-	[K] Transcription	-	COG1414
366bp	0.33	121aa	transcriptional regulator, GntR family	GntR [pfam00392]	-	-	-	[K] Transcription	-	COG1725
285bp	0.36	94aa	transcriptional regulator, XRE family	HTH_19 [pfam12844]	-	-	-	[K] Transcription	-	COG1813
291bp	0.37	96aa	Predicted transcriptional regulator	HTH_3 [pfam01381]	-	-	-	[K] Transcription	-	COG2944
264bp	0.34	87aa	Trp operon repressor	Trp_repressor [pfam01371]	-	K03720	TrpR family transcriptional regulator, trp operon repres	[K] Transcription	-	COG2973
2178bp	0.37	725aa	Transcriptional regulator containing GAF, AAA-type ATF	Sigma54_activat [pfam00158]	-	K15836	formate hydrogenlyase transcriptional activator0.00E+	[K] Transcription	-	COG3604
210bp	0.33	69aa	Predicted transcriptional regulator	HTH_26 [pfam13443]	-	K07727	putative transcriptional regulator2.00E-18	[K] Transcription	-	COG3655
285bp	0.42	95aa	Transposase and inactivated derivatives	DDE_Tnp_IS240 [pfam13610]	-	-	-	[L] Replication, recombination and repair	-	COG3316
360bp	0.38	119aa	Transposase and inactivated derivatives	TnpB_IS66 [pfam05717]	-	-	-	[L] Replication, recombination and repair	-	COG3436
570bp	0.33	189aa	restart primosome assembly protein PriC	priB_priC [pfam07445]	-	K04067	primosomal replication protein N''0.00E+00	[L] Replication, recombination and repair	-	COG3923
651bp	0.31	216aa	Glycosyltransferase involved in LPS biosynthesis	Glyco_transf_25 [pfam01755]	-	K07270	glycosyl transferase, family 258.70E-28	[M] Cell wall/membrane/envelope biogenesis	-	COG3306
522bp	0.4	173aa	Flagellar basal body-associated protein	FliL [pfam03748]	-	K02415	flagellar FliL protein2.60E-38	[N] Cell motility	-	COG1580
414bp	0.32	137aa	flagellar biosynthetic protein FliO	FliO [pfam04347]	-	K02418	flagellar protein FliO/FliZ1.00E-13	[N] Cell motility	-	COG3190
300bp	0.29	99aa	hypothetical protein	-	-	-	-	[P] Inorganic ion transport and metabolism	-	COG0748
600bp	0.39	199aa	Uncharacterized protein involved in copper resistance	CutC [pfam03932]	-	K06201	copper homeostasis protein0.00E+00	[P] Inorganic ion transport and metabolism	-	COG3142
2280bp	0.38	759aa	ferrous iron transporter FeoB	Gate [pfam07670]	-	K04759	ferrous iron transport protein B0.00E+00	[P] Inorganic ion transport and metabolism	-	COG0370
231bp	0.36	76aa	Fe2+ transport system protein A	FeoA [pfam04023]	-	K04758	ferrous iron transport protein A1.20E-11	[P] Inorganic ion transport and metabolism	-	COG1918
399bp	0.4	132aa	Predicted metal-dependent hydrolase with the TIM-bar	Amidohydro_3 [pfam07969]	-	-	-	[R] General function prediction only	-	COG1574
912bp	0.38	303aa	Predicted Fe-S oxidoreductase	Radical_SAM [pfam04055]	-	-	-	[R] General function prediction only	-	COG1242
2067bp	0.37	688aa	Esterase/lipase	Acyltransferase [pfam01553]	-	-	-	[R] General function prediction only	-	COG1647
648bp	0.35	215aa	ABC-type uncharacterized transport system, permease	NicO [pfam03824]	-	-	-	[R] General function prediction only	-	COG2215

Table S2.7 Unique genes and their respective COG, KEGG, EC and Pfam classifications continued

PRT Marinosulfonomonas									
DNA_length	**GC**	**Protein_length**	**Product_name**	**pfam**	**EC**	**KO_id**	**KO**	**COG_category**	**KEGG_module**	**COG_id**
1248bp	0.53	415aa	AAA domain.	AAA_31 [pfam13614]	-	K02282	pilus assembly protein CpaE0.00E+00	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG4963
822bp	0.57	273aa	Glycerol kinase	FGGY_C [pfam02782]	-	-	-	[C] Energy production and conversion	-	COG0554
1761bp	0.51	586aa	nitrous oxide reductase apoprotein	COX2 [pfam00116]	Nitrous-oxide reductase.	K00376	nitrous-oxide reductase [EC:1.7.2.4]0.00E+00	[C] Energy production and conversion	M00529: Denitrification, nitrate => nitrogen	COG4263
771bp	0.56	256aa	ABC-type dipeptide/oligopeptide/nickel transport syste	ABC_tran [pfam00005]	-	K02032	peptide/nickel transport system ATP-binding protein0.0	[E] Amino acid transport and metabolism	M00239: Peptides/nickel transport system	COG1124
405bp	0.55	134aa	methylmalonyl-CoA epimerase	Glyoxalase_4 [pfam13669]	Methylmalonyl-CoA epimerase.	K05606	methylmalonyl-CoA/ethylmalonyl-CoA epimerase [EC:5	[E] Amino acid transport and metabolism	M00375: Hydroxypropionate-hydroxybutylate cycle	COG3185
306bp	0.51	101aa	Tryptophan 2,3-dioxygenase (vermilion)	Trp_dioxygenase [pfam03301]	-	-	-	[E] Amino acid transport and metabolism	-	COG3483
837bp	0.49	278aa	Tryptophan 2,3-dioxygenase holoenzyme (EC 1.13.11.	Trp_dioxygenase [pfam03301]	Tryptophan 2,3-dioxygenase.	K00453	tryptophan 2,3-dioxygenase [EC:1.13.11.11]0.00E+00	[E] Amino acid transport and metabolism	M00038: Tryptophan metabolism,tryptophan=>kynure	COG3483
1038bp	0.59	345aa	Beta-glucosidase-related glycosidases	Glyco_hydro_3 [pfam00933]	Beta-N-acetylhexosaminidase.	K01207	beta-N-acetylhexosaminidase [EC:3.2.1.52]0.00E+00	[G] Carbohydrate transport and metabolism	-	COG1472
930bp	0.56	309aa	fructose-bisphosphate aldolase (EC 4.1.2.13)	DeoC [pfam01791]	Fructose-bisphosphate aldolase.	K01623	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]0.	[G] Carbohydrate transport and metabolism	M00165: Reductive pentose phosphate cycle (Calvin cy	COG1830
978bp	0.5	325aa	biotin synthase (EC 2.8.1.6)	BATS [pfam06968]	Biotin synthase.	K01012	biotin synthetase [EC:2.8.1.6]0.00E+00	[H] Coenzyme transport and metabolism	M00123: Biotin biosynthesis, pimeloyl-CoA => biotin	COG0502
1164bp	0.6	388aa	Cobalamin biosynthesis protein CbiG	CbiG_N [pfam11760]	-	-	-	[H] Coenzyme transport and metabolism	-	COG2073
492bp	0.56	163aa	cyclic pyranopterin monophosphate synthase subunit M	MoaC [pfam01967]	-	K03637	molybdenum cofactor biosynthesis protein C0.00E+00	[H] Coenzyme transport and metabolism	-	COG0315
846bp	0.57	282aa	FAD binding domain.	FAD_binding_2 [pfam00890]	-	-	-	[H] Coenzyme transport and metabolism	-	COG0029
723bp	0.56	240aa	FAD binding domain./Fumarate reductase flavoprotein	FAD_binding_2 [pfam00890]	-	-	-	[H] Coenzyme transport and metabolism	-	COG0029
444bp	0.57	147aa	molybdopterin synthase subunit MoaE (EC 2.8.1.12)	MoaE [pfam02391]	Transferases.	K03635	molybdopterin synthase catalytic subunit [EC:2.-.-.-]0.	[H] Coenzyme transport and metabolism	-	COG0314
852bp	0.53	283aa	nicotinate-nucleotide pyrophosphorylase [carboxylating	QRPTase_N [pfam02749]	Nicotinate-nucleotide diphosphorylase (carboxylating).	K00767	nicotinate-nucleotide pyrophosphorylase (carboxylating	[H] Coenzyme transport and metabolism	M00115: NAD biosynthesis, aspartate => NAD	COG0157
843bp	0.54	280aa	pantothenate synthetase (EC 6.3.2.1)	Pantoate_ligase [pfam02569]	Pantoate--beta-alanine ligase.	K01918	pantoate--beta-alanine ligase [EC:6.3.2.1]0.00E+00	[H] Coenzyme transport and metabolism	M00119: Pantothenate biosynthesis, valine/L-aspartate	COG0414
621bp	0.5	207aa	Quinolinate synthase	NadA [pfam02445]	-	-	-	[H] Coenzyme transport and metabolism	-	COG0379
1092bp	0.53	363aa	thiamine biosynthesis protein ThiC	ThiC [pfam01964]	-	-	-	[H] Coenzyme transport and metabolism	-	COG0422
210bp	0.29	70aa	ABC-type long-chain fatty acid transport system, fused	SbmA_BacA [pfam05992]	-	-	-	[I] Lipid transport and metabolism	-	COG1133
1065bp	0.48	354aa	Long-chain fatty acid transport protein	Toluene_X [pfam03349]	-	-	-	[I] Lipid transport and metabolism	-	COG2067
1512bp	0.55	503aa	aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase sub	GatB_Yqey [pfam02637]	Asparaginyl-tRNA synthase (glutamine-hydrolyzing).	K02434	aspartyl-tRNA(Asn)/glutamyl-tRNA (Gln) amidotransfer	[J] Translation, ribosomal structure and biogenesis	-	COG0064
135bp	0.54	44aa	LSU ribosomal protein L34P	Ribosomal_L34 [pfam00468]	-	K02914	large subunit ribosomal protein L34.90E-15	[J] Translation, ribosomal structure and biogenesis	M00178: Ribosome, bacteria	COG0230
564bp	0.52	187aa	translation elongation factor P (EF-P)	EFP [pfam01132]	-	K02356	elongation factor P0.00E+00	[J] Translation, ribosomal structure and biogenesis	-	COG0231
597bp	0.54	198aa	Domain of unknown function (DUF4095).	DUF4095 [pfam13338]	-	-	-	[K] Transcription	-	COG5340
261bp	0.48	86aa	Predicted transcriptional regulator	Peptidase_S24 [pfam00717]	-	-	-	[K] Transcription	-	COG2932
663bp	0.51	220aa	Predicted transcriptional regulator	Peptidase_S24 [pfam00717]	-	-	-	[K] Transcription	-	COG2932
657bp	0.47	218aa	Predicted transcriptional regulator	HTH_3 [pfam01381]	-	-	-	[K] Transcription	-	COG2932
1644bp	0.54	547aa	Regulator of nitric oxide reductase transcription	FMN_bind [pfam04205]	-	-	-	[K] Transcription	-	COG3901
282bp	0.57	94aa	addiction module antitoxin, RelB/DinJ family	RelB [pfam04221]	-	K07473	DNA-damage-inducible protein J3.50E-29	[L] Replication, recombination and repair	-	COG3077
459bp	0.57	152aa	DNA polymerase III, chi subunit (EC 2.7.7.7)	DNA_pol3_chi [pfam04364]	DNA-directed DNA polymerase.	K02339	DNA polymerase III subunit chi [EC:2.7.7.7]0.00E+00	[L] Replication, recombination and repair	M00260: DNA polymerase III complex, bacteria	COG2927
960bp	0.45	319aa	Plasmid rolling circle replication initiator protein and tr	Rep_1 [pfam01446]	-	-	-	[L] Replication, recombination and repair	-	COG5655
816bp	0.51	271aa	Site-specific DNA methylase	Methyltransf012 [pfam02086]	Site-specific DNA-methyltransferase (adenine-specific).	K06223	DNA adenine methylase [EC:2.1.1.72]0.00E+00	[L] Replication, recombination and repair	-	COG0338
438bp	0.5	145aa	Glycerol-3-phosphate cytidylyltransferase (EC 2.7.7.39	CTP_transf_2 [pfam01467]	Glycerol-3-phosphate cytidylyltransferase.	K00980	glycerol-3-phosphate cytidylyltransferase [EC:2.7.7.39	[M] Cell wall/membrane/envelope biogenesis	-	COG2870
3882bp	0.51	1293aa	Rhs family protein	RHS_repeat [pfam05593]	-	-	-	[M] Cell wall/membrane/envelope biogenesis	-	COG3209
3882bp	0.51	1293aa	RHS repeat-associated core domain	RHS_repeat [pfam05593]	-	-	-	[M] Cell wall/membrane/envelope biogenesis	-	COG3209
480bp	0.47	159aa	Chemotaxis protein; stimulates methylation of MCP prc	CheD [pfam03975]	Protein-glutamine glutaminase.	K03411	chemotaxis protein CheD [EC:3.5.1.44]4.80E-39	[N] Cell motility	-	COG1871
984bp	0.52	327aa	type II secretion system protein F (GspF)	T2SF [pfam00482]	-	K12511	tight adherence protein C0.00E+00	[N] Cell motility	-	COG2064
1161bp	0.56	386aa	Hydrogenase/urease accessory protein	HupE_UreJ_2 [pfam13795]	-	-	-	[O] Posttranslational modification, protein turnover, ch	-	COG2370
489bp	0.56	162aa	methionine-R-sulfoxide reductase	SelR [pfam01641]	Peptide-methionine (R)-S-oxide reductase.	K07305	peptide-methionine (R)-S-oxide reductase [EC:1.8.4.1:	[O] Posttranslational modification, protein turnover, ch	-	COG0229
903bp	0.6	301aa	Cyanate permease	MFS_1 [pfam07690]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG2807
792bp	0.56	263aa	Domain of unknown function (DUF4198).	DUF4198 [pfam10670]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG5266
441bp	0.56	146aa	multisubunit sodium/proton antiporter, MrpB subunit (1	MnhB [pfam04039]	-	K05566	multicomponent Na+:H+ antiporter subunit B1.90E-42	[P] Inorganic ion transport and metabolism	-	COG2111
1698bp	0.51	565aa	Na+/phosphate symporter	Na_Pi_cotrans [pfam02690]	-	K03324	phosphate:Na+ symporter0.00E+00	[P] Inorganic ion transport and metabolism	-	COG1283
354bp	0.56	117aa	Nitrous oxidase accessory protein	NosD [pfam05048]	-	-	-	[P] Inorganic ion transport and metabolism	-	COG3420
567bp	0.57	188aa	Predicted subunit of the Multisubunit Na+/H+ antiporte	DUF4040 [pfam13244]	-	K05566	multicomponent Na+:H+ antiporter subunit B0.00E+00	[P] Inorganic ion transport and metabolism	-	COG1563
963bp	0.58	320aa	Tellurite resistance protein and related permeases	C4dic_mal_tran [pfam03595]	-	K03304	tellurite resistance/dicarboxylate transporter, TDT fami	[P] Inorganic ion transport and metabolism	-	COG1275
1146bp	0.57	381aa	conserved hypothetical protein, YceG family	YceG [pfam02618]	-	K07082	UPF0755 protein0.00E+00	[R] General function prediction only	-	COG1559
435bp	0.56	144aa	FOG: CBS domain	CBS [pfam00571]	-	-	-	[R] General function prediction only	-	COG0517
525bp	0.58	174aa	folate-binding protein YgfZ	GCV_T_C [pfam08669]	-	-	-	[R] General function prediction only	-	COG0354
663bp	0.57	220aa	Methyltransferase domain.	Methyltransf_11 [pfam08241]	-	-	-	[R] General function prediction only	-	COG4976
1245bp	0.53	414aa	phosphonoacetate hydrolase (EC 3.11.1.2)	Phosphodiest [pfam01663]	-	-	-	[R] General function prediction only	-	COG1524
414bp	0.53	137aa	Predicted acyltransferase	Acetyltransf_7 [pfam13508]	-	K02348	ElaA protein2.90E-37	[R] General function prediction only	-	COG2153
411bp	0.55	136aa	Predicted dioxygenase of extradiol dioxygenase family	Glyoxalase_2 [pfam12681]	-	-	-	[R] General function prediction only	-	COG3565
927bp	0.56	308aa	Predicted metal-dependent phosphoesterases (PHP fam	PHP [pfam02811]	-	-	-	[R] General function prediction only	-	COG0613
924bp	0.58	307aa	Predicted metal-dependent phosphoesterases (PHP fam	PHP [pfam02811]	-	-	-	[R] General function prediction only	-	COG0613
549bp	0.6	182aa	Predicted nucleic acid-binding protein, contains PIN dor	PIN_3 [pfam13470]	-	-	-	[R] General function prediction only	-	COG1569
921bp	0.58	306aa	Predicted permease, DMT superfamily	EamA [pfam00892]	-	-	-	[R] General function prediction only	-	COG5006
1059bp	0.57	352aa	Predicted permeases	DUF318 [pfam03773]	-	-	-	[R] General function prediction only	-	COG0701
672bp	0.61	223aa	Protein of unknown function (DUF938).	DUF938 [pfam06080]	-	-	-	[R] General function prediction only	-	COG0500
1251bp	0.39	416aa	tape measure domain	-	-	-	-	[R] General function prediction only	-	COG3941
1053bp	0.59	350aa	Predicted integral membrane protein	UPF0104 [pfam03706]	-	-	-	[S] Function unknown	-	COG0392
471bp	0.57	156aa	Predicted integral membrane protein	Metal_resist [pfam13801]	-	-	-	[S] Function unknown	-	COG5612
906bp	0.57	301aa	Predicted membrane protein	EamA [pfam00892]	-	-	-	[S] Function unknown	-	COG2510
498bp	0.58	165aa	Predicted membrane protein	DUF2306 [pfam10067]	-	-	-	[S] Function unknown	-	COG5395
489bp	0.59	162aa	Predicted secreted (periplasmic) protein	LptE [pfam04390]	-	K03643	LPS-assembly lipoprotein1.80E-41	[S] Function unknown	-	COG5468
273bp	0.52	90aa	Predicted small integral membrane protein	DUF2160 [pfam09928]	-	-	-	[S] Function unknown	-	COG5477
459bp	0.53	152aa	Protein of unknown function (DUF1499).	DUF1499 [pfam07386]	-	-	-	[S] Function unknown	-	COG4446
519bp	0.59	172aa	EF hand./EF-hand domain pair.	EF_hand_5 [pfam13499]	-	-	-	[T] Signal transduction mechanisms	-	COG5126
1434bp	0.57	478aa	His Kinase A (phospho-acceptor) domain./Histidine kin	HATPase_c [pfam02518]	-	-	-	[T] Signal transduction mechanisms	-	COG2205
885bp	0.56	294aa	Mn2+-dependent serine/threonine protein kinase	Kdo [pfam06293]	-	-	-	[T] Signal transduction mechanisms	-	COG3642
507bp	0.58	168aa	Phosphohistidine phosphatase SixA	His_Phos_1 [pfam00300]	Hydrolases. Acting on ester bonds. Phosphoric monoest	K08296	phosphohistidine phosphatase [EC:3.1.3.-]8.60E-39	[T] Signal transduction mechanisms	-	COG2062
1137bp	0.55	378aa	Signal transduction histidine kinase, nitrogen specific	HisKA [pfam00512]	Histidine kinase.	K07708	two-component system, NtrC family, nitrogen regulatio	[T] Signal transduction mechanisms	M00497: GlnL-GlnG (nitrogen regulation) two-compone	COG3852
897bp	0.53	298aa	Flp pilus assembly protein CpaB	-	-	K02279	pilus assembly protein CpaB0.00E+00	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG3745
1710bp	0.56	569aa	Flp pilus assembly protein TadD, contains TPR repeats	TPR_11 [pfam13414]	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG5010
501bp	0.51	166aa	Flp pilus assembly protein, protease CpaA	Peptidase_A24 [pfam01478]	Prepilin peptidase.	K02278	prepilin peptidase CpaA [EC:3.4.23.43]2.80E-42	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG4960
1434bp	0.55	477aa	Flp pilus assembly protein, secretin CpaC	BON [pfam04972]	-	K02280	pilus assembly protein CpaC0.00E+00	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG4964
186bp	0.52	61aa	protein translocase subunit secE/sec61 gamma	SecE [pfam00584]	-	K03073	preprotein translocase subunit SecE5.00E-21	[U] Intracellular trafficking, secretion, and vesicular tra	M00335: Sec (secretion) system	COG0690
966bp	0.56	321aa	protein translocase subunit secF	SecD_SecF [pfam02355]	-	K03074	preprotein translocase subunit SecF0.00E+00	[U] Intracellular trafficking, secretion, and vesicular tra	M00335: Sec (secretion) system	COG0341
1602bp	0.49	533aa	Type II secretory pathway, ATPase PulE/Tfp pilus asser	T2SE_Nter [pfam05157]	-	K02454	general secretion pathway protein E0.00E+00	[U] Intracellular trafficking, secretion, and vesicular tra	M00331: Type II general secretion system	COG2804

Table S2.7 Unique genes and their respective COG, KEGG, EC and Pfam classifications continued

PRT Marinosulfonomonas DNA_length	GC	Protein_length	Product_name	pfam	EC	KO_id	KO	COG_category	KEGG_module	COG_id
723bp	0.59	240aa	ABC-type multidrug transport system, permease comp	ABC2_membrane_3 [pfam12698]	-	-	-	[V] Defense mechanisms	-	COG0842
276bp	0.48	91aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
276bp	0.48	91aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
339bp	0.54	112aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
276bp	0.48	91aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
279bp	0.51	92aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	K07483	transposase3.60E-23	[L] Replication, recombination and repair	-	COG2963
339bp	0.54	112aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
276bp	0.48	91aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
279bp	0.51	92aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	K07483	transposase3.60E-23	[L] Replication, recombination and repair	-	COG2963
276bp	0.48	91aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
339bp	0.54	112aa	Transposase and inactivated derivatives	HTH_Tnp_1 [pfam01527]	-	-	-	[L] Replication, recombination and repair	-	COG2963
780bp	0.48	259aa	hypothetical protein	-	-	-	-	[V] Defense mechanisms	-	COG1401
234bp	0.38	78aa	hypothetical protein	-	-	-	-	[K] Transcription	-	COG2002
858bp	0.56	285aa	hypothetical protein	-	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG5010
570bp	0.54	189aa	hypothetical protein	-	-	-	-	[U] Intracellular trafficking, secretion, and vesicular tra	-	COG5010
2103bp	0.56	700aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG5283
231bp	0.61	76aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG5457
1572bp	0.54	523aa	hypothetical protein	-	-	-	-	[S] Function unknown	-	COG5617
660bp	0.51	219aa	Uncharacterized conserved protein	YceI [pfam04264]	-	-	-	[S] Function unknown	-	COG2353
576bp	0.54	191aa	Uncharacterized conserved protein	YceI [pfam04264]	-	-	-	[S] Function unknown	-	COG2353
1806bp	0.58	601aa	Uncharacterized conserved protein	DUF2156 [pfam09924]	-	-	-	[S] Function unknown	-	COG2898
1788bp	0.59	595aa	Uncharacterized conserved protein	UPF0104 [pfam03706]	-	-	-	[S] Function unknown	-	COG2898
504bp	0.55	167aa	Uncharacterized conserved protein	FIIB [pfam03692]	-	K09160	hypothetical protein0.00E+00	[S] Function unknown	-	COG2983
675bp	0.59	224aa	Uncharacterized conserved protein	DUF1028 [pfam06267]	-	-	-	[S] Function unknown	-	COG3342
387bp	0.55	128aa	Uncharacterized conserved protein	DUF302 [pfam03625]	-	-	-	[S] Function unknown	-	COG3439
735bp	0.53	245aa	Uncharacterized conserved protein	DUF297 [pfam03537]	-	-	-	[S] Function unknown	-	COG3868
837bp	0.49	278aa	Uncharacterized conserved protein	DUF1445 [pfam07286]	-	-	-	[S] Function unknown	-	COG4336
1086bp	0.56	361aa	Uncharacterized conserved protein, contains double-str	Cupin_2 [pfam07883]	-	-	-	[S] Function unknown	-	COG1917
486bp	0.56	161aa	Uncharacterized conserved protein, contains double-str	Cupin_2 [pfam07883]	-	-	-	[S] Function unknown	-	COG3837
2127bp	0.59	708aa	Uncharacterized membrane protein	DUF1355 [pfam07090]	-	-	-	[S] Function unknown	-	COG5426
1233bp	0.56	410aa	Uncharacterized protein conserved in bacteria	RmuC [pfam02646]	-	K09760	DNA recombination protein RmuC0.00E+00	[S] Function unknown	-	COG1322
621bp	0.51	206aa	Uncharacterized protein conserved in bacteria	DUF452 [pfam04301]	-	K09789	hypothetical protein2.70E-21	[S] Function unknown	-	COG2830
354bp	0.48	117aa	Uncharacterized protein conserved in bacteria	DUF454 [pfam04304]	-	K09790	hypothetical protein1.70E-33	[S] Function unknown	-	COG2832
708bp	0.55	235aa	Uncharacterized protein conserved in bacteria	DUF484 [pfam04340]	-	K09921	hypothetical protein0.00E+00	[S] Function unknown	-	COG3159
588bp	0.57	195aa	Uncharacterized protein conserved in bacteria	DUF2064 [pfam09837]	-	K09931	hypothetical protein0.00E+00	[S] Function unknown	-	COG3222
339bp	0.55	112aa	Uncharacterized protein conserved in bacteria	DUF952 [pfam06108]	-	-	-	[S] Function unknown	-	COG3502
528bp	0.55	175aa	Uncharacterized protein conserved in bacteria	DUF1287 [pfam06940]	-	K09974	hypothetical protein0.00E+00	[S] Function unknown	-	COG3738
1122bp	0.55	373aa	Uncharacterized protein conserved in bacteria	Prim_Zn_Ribbon [pfam08273]	-	-	-	[S] Function unknown	-	COG4643
282bp	0.45	93aa	Uncharacterized protein conserved in bacteria	-	-	-	-	[S] Function unknown	-	COG5304
330bp	0.55	109aa	Uncharacterized protein conserved in bacteria	ATPase_gene1 [pfam09527]	-	K02116	ATP synthase protein I2.00E-28	[S] Function unknown	-	COG5336
3363bp	0.57	1120aa	Uncharacterized protein involved in outer membrane bi	AsmA_2 [pfam13502]	-	-	-	[M] Cell wall/membrane/envelope biogenesis	-	COG2982
228bp	0.55	75aa	Uncharacterized small protein	DUF1150 [pfam06620]	-	-	-	[S] Function unknown	-	COG5568
165bp	0.44	54aa	Uncharacterized small protein	DUF465 [pfam04325]	-	-	-	[S] Function unknown	-	COG5570
537bp	0.52	178aa	phage baseplate assembly protein V	Phage_base_V [pfam04717]	-	-	-	[R] General function prediction only	-	COG4540
339bp	0.57	112aa	Phage baseplate assembly protein W	GPW_gp25 [pfam04965]	-	-	-	[R] General function prediction only	-	COG3628
390bp	0.53	129aa	Phage baseplate assembly protein W	GPW_gp25 [pfam04965]	-	-	-	[R] General function prediction only	-	COG3628
336bp	0.52	111aa	Phage baseplate assembly protein W	GPW_gp25 [pfam04965]	-	-	-	[R] General function prediction only	-	COG3628
507bp	0.55	168aa	phage contractile tail tube protein, P2 family	Phage_tube [pfam04985]	-	-	-	[R] General function prediction only	-	COG3498
339bp	0.54	112aa	phage head-tail adaptor, putative, SPP1 family	Phage_H_T_join [pfam05521]	-	-	-	[R] General function prediction only	-	COG5614
471bp	0.55	156aa	Phage P2 baseplate assembly protein gpV	Phage_base_V [pfam04717]	-	-	-	[R] General function prediction only	-	COG4540
318bp	0.55	105aa	Phage P2 baseplate assembly protein gpV	Phage_base_V [pfam04717]	-	-	-	[R] General function prediction only	-	COG4540
993bp	0.52	330aa	Phage protein D	Phage_GPD [pfam05954]	-	-	-	[R] General function prediction only	-	COG3500
855bp	0.57	284aa	Phage protein D	Phage_GPD [pfam05954]	-	-	-	[R] General function prediction only	-	COG3500
411bp	0.54	136aa	Phage protein U	Phage_P2_GpU [pfam06995]	-	-	-	[R] General function prediction only	-	COG3499
423bp	0.52	140aa	Phage protein U	Phage_P2_GpU [pfam06995]	-	-	-	[R] General function prediction only	-	COG3499
642bp	0.52	213aa	phage tail protein, P2 protein I family	Tail_P2_I [pfam09684]	-	-	-	[R] General function prediction only	-	COG4385
570bp	0.57	189aa	phage tail protein, P2 protein I family	Tail_P2_I [pfam09684]	-	-	-	[R] General function prediction only	-	COG4385
507bp	0.51	168aa	Phage tail tube protein FII	Phage_tube [pfam04985]	-	-	-	[R] General function prediction only	-	COG3498
441bp	0.5	146aa	Phage terminase, small subunit	Terminase_4 [pfam05119]	-	-	-	[L] Replication, recombination and repair	-	COG3747
1668bp	0.51	555aa	Phage terminase-like protein, large subunit	Terminase_1 [pfam03354]	-	-	-	[R] General function prediction only	-	COG4626
480bp	0.49	159aa	phage virion morphogenesis (putative tail completion)	Phage_tail_S [pfam05069]	-	-	-	[R] General function prediction only	-	COG5005
1653bp	0.53	550aa	phage/plasmid primase, P4 family, C-terminal domain	D5_N [pfam08706]	-	K06919	putative DNA primase/helicase0.00E+00	[R] General function prediction only	-	COG3378
903bp	0.61	300aa	Phage-related baseplate assembly protein	Baseplate_J [pfam04865]	-	-	-	[R] General function prediction only	-	COG3948
639bp	0.49	213aa	Phage-related baseplate assembly protein	Baseplate_J [pfam04865]	-	-	-	[R] General function prediction only	-	COG3948
957bp	0.54	318aa	Phage-related baseplate assembly protein	Baseplate_J [pfam04865]	-	-	-	[R] General function prediction only	-	COG3948
759bp	0.56	252aa	Phage-related lysozyme (muraminidase)	Phage_lysozyme [pfam00959]	-	-	-	[R] General function prediction only	-	COG3772
735bp	0.56	244aa	Phage-related lysozyme (muraminidase)	Phage_lysozyme [pfam00959]	-	-	-	[R] General function prediction only	-	COG3772
2724bp	0.54	907aa	Phage-related minor tail protein.	PhageMin_Tail [pfam10145]	-	-	-	[S] Function unknown	-	COG5283
210bp	0.53	69aa	P2-like prophage tail protein X	Phage_tail_X [pfam05489]	-	-	-	[R] General function prediction only	-	COG5004
225bp	0.56	74aa	P2-like prophage tail protein X	Phage_tail_X [pfam05489]	-	-	-	[R] General function prediction only	-	COG5004
1644bp	0.51	547aa	Mu-like prophage FluMu protein gp28	Terminase_6 [pfam03237]	-	-	-	[R] General function prediction only	-	COG4373
504bp	0.52	167aa	Mu-like prophage I protein.	Mu-like_Pro [pfam10123]	-	-	-	[R] General function prediction only	-	COG4388
594bp	0.53	198aa	Mu-like prophage I protein.	Mu-like_Pro [pfam10123]	-	-	-	[R] General function prediction only	-	COG4388
393bp	0.51	130aa	Mu-like prophage major head subunit gpT.	Mu-like_gpT [pfam10124]	-	-	-	[R] General function prediction only	-	COG4397
504bp	0.48	168aa	Mu-like prophage major head subunit gpT.	Mu-like_gpT [pfam10124]	-	-	-	[R] General function prediction only	-	COG4397
447bp	0.58	148aa	Mu-like prophage protein gp16	DUF1018 [pfam06252]	-	-	-	[S] Function unknown	-	COG4382
648bp	0.51	215aa	Mu-like prophage protein gp16	DUF1018 [pfam06252]	-	-	-	[S] Function unknown	-	COG4382
1617bp	0.52	538aa	Mu-like prophage protein gp29	DUF935 [pfam06074]	-	-	-	[S] Function unknown	-	COG4383
312bp	0.56	103aa	Mu-like prophage protein gp36	DUF1320 [pfam07030]	-	-	-	[S] Function unknown	-	COG4387
1017bp	0.56	338aa	Bacteriophage protein gp37	Gp37_Gp68 [pfam07505]	-	-	-	[S] Function unknown	-	COG4422
690bp	0.53	229aa	Uncharacterized ATPase, putative transposase	Phage-MuB_C [pfam09077]	-	-	-	[R] General function prediction only	-	COG2842
1242bp	0.53	413aa	Uncharacterized protein, homolog of phage Mu protein	Phage_Mu_F [pfam04233]	-	-	-	[S] Function unknown	-	COG2369
267bp	0.49	88aa	Uncharacterized protein conserved in bacteria	DUF497 [pfam04365]	-	K09803	hypothetical protein1.00E-14	[S] Function unknown	-	COG2929

REFERENCES

Allen EE, Bartlett DH (2002). Structure and regulation of the omega-3 polyunsaturated fatty acid synthase genes from the deep-sea bacterium Photobacterium profundum strain SS9. *Microbiology* **148:** 1903-1913.

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

Arakawa S, Sato T, Sato R, Zhang J, Gamo T, Tsunogai U, Hirota A, Yoshida Y, Usami R, Inagaki F, Kato C (2006). Molecular phylogenetic and chemical analyses of the microbial mats in deep-sea cold seep sediments at the northeastern Japan Sea. *Extremophiles* **10:** 311-319.

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A Lesin, VM, Nikolenko, SI, Pham, S, Prjibelski, AD, Pyshkin, AV, Sirotkin, AV, Vyahhi, N, Tesler, G, Alekseyev, MA, Pevzner, PA (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19:** 455-477.

Bartlett DH (2002). Pressure effects on in vivo microbial processes. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1595:** 367-381.

Blankenship L, Yayanos A, Cadien D, Levin L (2006). Vertical zonation patterns of scavenging amphipods from the Hadal zone of the Tonga and Kermadec Trenches. *Deep-Sea Research Part I-Oceanographic Research Papers* **53:** 48-61.

Borziak K, Posner M, Upadhyay A, Danson M, Bagby S, Dorus S (2014). Comparative Genomic Analysis Reveals 2-Oxoacid Dehydrogenase Complex Lipoylation Correlation with Aerobiosis in Archaea. *Plos One* **9**.

Calamita G (2000). The Escherichia coli aquaporin-Z water channel. *Molecular Microbiology* **37:** 254-262.

Campanaro S, De Pascale F, Telatin A, Schiavon R, Bartlett D, Valle G (2012). The transcriptional landscape of the deep-sea bacterium Photobacterium profundum in both a toxR mutant and its parental strain. *BMC Genomics* **13:** 567.

Carini P, Steindler L, Beszteri S, Giovannoni S (2013). Nutrient requirements for growth of the extreme oligotroph 'Candidatus Pelagibacter ubique' HTCC1062 on a defined medium. *Isme Journal* **7:** 592-602.

Carty S, Sreekumar K, Raetz C (1999). Effect of cold shock on lipid A biosynthesis in Escherichia coli - Induction at 12 degrees C of an acyltransferase specific for

palmitoleoyl-acyl carrier protein. *Journal of Biological Chemistry* **274:** 9677-9685.

DeLong E, Franks D, Yayanos A (1997). Evolutionary relationships of cultivated psychrophilic and barophilic deep-sea bacteria. *Applied and Environmental Microbiology* **63:** 2105-2108.

DeLong E, Preston C, Mincer T, Rich V, Hallam S, Frigaard N Martinez, A Sullivan, MB, Edwards, R, Brito, BR, Chisholm, SW, Karl, DM (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496-503.

Durbin AM, Teske A (2010). Sediment-associated microdiversity within the Marine Group I Crenarchaeota. *Environmental Microbiology Reports* **2:** 693-703.

Eichhorn E, van der Ploeg JR, Kertesz MA, Leisinger T (1997). Characterization of α-Ketoglutarate-dependent Taurine Dioxygenase from Escherichia coli. *Journal of Biological Chemistry* **272:** 23031-23036.

Eloe E, Fadrosh D, Novotny M, Allen L, Kim M, Lombardo M Yee-Greenbaum, J, Yooseph, S, Allen, EE, Lasken, R, Williamson, SJ, Bartlett, DH (2011a). Going Deeper: Metagenome of a Hadopelagic Microbial Community. *Plos One* **6**.

Eloe E, Malfatti F, Gutierrez J, Hardy K, Schmidt W, Pogliano K, Pogliano, J Azam, F, Bartlett, DH (2011b). Isolation and Characterization of a Psychropiezophilic Alphaproteobacterium. *Applied and Environmental Microbiology* **77:** 8145-8153.

Eloe E, Shulse C, Fadrosh D, Williamson S, Allen E, Bartlett D (2011c). Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports* **3:** 449-458.

Fang J, Zhang L, Bazylinski DA (2010). Deep-sea piezosphere and piezophiles: geomicrobiology and biogeochemistry. *Trends in Microbiology* **18:** 413-422.

Fischer S, Maier L, Stoll B, Brendel J, Fischer E, Pfeiffer F, Dyall-Smith, M Marchfelder, A (2012). An Archaeal Immune System Can Detect Multiple Protospacer Adjacent Motifs (PAMs) to Target Invader DNA. *Journal of Biological Chemistry* **287:** 33351-33363.

France S (1993). Geographic-variation among 3 isolated populations of the hadal amphipod Hirondellea gigas (Crustacea, Amphipoda, Lysianassoidea). *Marine Ecology Progress Series* **92:** 277-287.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean.

Proceedings of the National Academy of Sciences of the United States of America **102:** 14683-14688.

Glud RN, Wenzhofer F, Middelboe M, Oguri K, Turnewitsch R, Canfield DE, Kitazato, H (2013). High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature Geosci* **6:** 284-288.

Goffredi S, Orphan V, Rouse G, Jahnke L, Embaye T, Turk K Lee, R Vrijenhoek, RC (2005). Evolutionary innovation: a bone-eating marine symbiosis. *Environmental Microbiology* **7:** 1369-1378.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé, MS. (2012). Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* **3**.

Herndl GJ, Reinthaler T, Teira E, van Aken H, Veth C, Pernthaler A Pernthaler, J (2005). Contribution of Archaea to Total Prokaryotic Production in the Deep Atlantic Ocean. *Applied and Environmental Microbiology* **71:** 2303-2309.

Hugler M, Gartner A, Imhoff J (2010). Functional genes as markers for sulfur cycling and CO2 fixation in microbial communities of hydrothermal vents of the Logatchev field. *Fems Microbiology Ecology* **73:** 526-537.

Inagaki F, Sakihama Y, Inoue A, Kato C, Horikoshi K (2002). Molecular phylogenetic analyses of reverse-transcribed bacterial rRNA obtained from deep-sea cold seep sediments. *Environmental Microbiology* **4:** 277-286.

Jamieson AJ (2001). Ecology of Deep Oceans: Hadal Trenches. *eLS*. John Wiley & Sons, Ltd.

Jones AC, Monroe EA, Podell S, Hess WR, Klages S, Esquenazi E, Niessen, S., Hoover, H, Rothmann, M, Lasken, RS, Yates, JR, Reinhardt, R, Kube, M, Burkart, MD, Allen, EE, Dorrestein, PC, Gerwick, WH, Gerwick, Lena (2011). Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium Lyngbya majuscula. *Proceedings of the National Academy of Sciences* **108:** 8815-8820.

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

Kim J, Jo B, Cha H (2011). Production of biohydrogen by heterologous expression of oxygen-tolerant Hydrogenovibrio marinus [NiFe]-hydrogenase in Escherichia coli. *Journal of Biotechnology* **155:** 312-319.

Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009). Comparative Metagenomic Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station

ALOHA in the North Pacific Subtropical Gyre. *Applied and Environmental Microbiology* **75:** 5345-5355.

Kumar M, Grzelakowski M, Zilles J, Clark M, Meier W (2007). Highly permeable polymeric membranes based on the incorporation of the functional water channel protein Aquaporin Z. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 20719-20724.

Kurz M, Brünig ANS, Galinski EA (2006). NhaD type sodium/proton-antiporter of Halomonas elongata: a salt stress response mechanism in marine habitats? *Saline Systems* **2:** 1-12.

Lai C-Y, Cronan JE (2003). β-Ketoacyl-Acyl Carrier Protein Synthase III (FabH) Is Essential for Bacterial Fatty Acid Synthesis. *Journal of Biological Chemistry* **278:** 51494-51503.

Lasken, RS (2012). *Genomic sequencing of uncultured microorganisms from single cells,* Nature Reviews Microbiology, 10: 631-640.

Lasken, RS and McLean, JS. Nature Reviews Genetics, in press

Lauro FM, Bartlett DH (2008). Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12:** 15-25.

Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere, M Z, Ting, L, Ertan, H, Johnson, J, Ferriera, S, Lapidus, A, Anderson, I, Kyrpides, N, Munk, AC, Detter, C, Han, CS, Brown, MV, Robb, FT, Kjelleberg, S, Cavicchioli, R (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci U S A* **106:** 15527-15533.

Lauro FM, Stratton TK, Chastain RA, Ferriera S, Johnson J, Goldberg SM Yayanos, AA, Bartlett, DH (2013). Complete Genome Sequence of the Deep-Sea Bacterium Psychromonas Strain CNPT3. *Genome Announc* **1**.

Le Bihan T, Rayner J, Roy MM, Spagnolo L (2013). <italic>Photobacterium profundum</italic> under Pressure: A MS-Based Label-Free Quantitative Proteomics Study. *PLoS ONE* **8:** e60897.

Leigh J, Dodsworth J (2007). Nitrogen regulation in bacteria and archaea. *Annual Review of Microbiology* **61:** 349-377.

Lokanath N, Kuroishi C, Okazaki N, Kunishima N (2004). Purification, crystallization and preliminary crystallographic analysis of the glycine-cleavage system component T-protein from Pyrococcus horikoshii OT3. *Acta Crystallographica Section D-Biological Crystallography* **60:** 1450-1452.

Lombard J, López-García P, Moreira D (2012). An ACP-Independent Fatty Acid Synthesis Pathway in Archaea: Implications for the Origin of Phospholipids. *Molecular Biology and Evolution* **29:** 3261-3265.

Lu L, Han W, Zhang J, Wu Y, Wang B, Lin X, Zhu, JG, Cai, ZC, Jia, ZJ (2012). Nitrification of archaeal ammonia oxidizers in acid soils is supported by hydrolysis of urea. *Isme Journal* **6:** 1978-1984.

Markowitz V, Chen I, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner, A Huang, JH, Woyke, T, Huntemann, M, Anderson, I, Billis, K, Varghese, N, Mavromatis, K, Pati, A, Ivanova, NN, Kyrpides, NC (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42:** D560-D567.

Martin D, Bartlett D, Roberts M (2002). Solute accumulation in the deep-sea bacterium Photobacterium profundum. *Extremophiles* **6:** 507-514.

Martin-Cuadrado A, Lopez-Garcia P, Alba J, Moreira D, Monticelli L, Strittmatter A, Gottschalk G, Rodriguez-Valera F (2007). Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat. *Plos One* **2**.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi, N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences* **110:** E2390-E2399.

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

Morris TW, Reed KE, Cronan JE (1994). Identification of the gene encoding lipoate-protein ligase A of Escherichia coli. Molecular cloning and characterization of the lplA gene and gene product. *Journal of Biological Chemistry* **269:** 16091-16100.

Muto A, Fujihara A, Ito K-i, Matsuno J, Ushida C, Himeno H (2000). Requirement of transfer-messenger RNA for the growth of Bacillus subtilis under stresses. *Genes to Cells* **5:** 627-635.

Niu B, Zhu Z, Fu L, Wu S, Li W (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27:** 1704-1705.

Pester M, Schleper C, Wagner M (2011). The Thaumarchaeota: an emerging view of

their phylogeny and ecophysiology. *Current Opinion in Microbiology* **14:** 300-306.

Podell S, Gaasterland T (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology* **8**.

Podell S, Ugalde J, Narasingarao P, Banfield J, Heidelberg K, Allen E (2013). Assembly- Driven Community Genomics of a Hypersaline Microbial Ecosystem. *Plos One* **8**.

Posner MG, Upadhyay A, Crennell SJ, Watson AJA, Dorus S, Danson MJ, Bagby, S (2013). Post-translational modification in the archaea: structural characterization of multi-enzyme complex lipoylation. *Biochemical Journal* **449:** 415-425.

Price M, Dehal P, Arkin A (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26:** 1641-1650.

Pruesse E, Peplies J, Glockner F (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28:** 1823-1829.

Rasmussen J, Vegge C, Frokiaer H, Howlett R, Krogfelt K, Kelly D, Ingmer, H (2013). Campylobacter jejuni carbon starvation protein A (CstA) is involved in peptide utilization, motility and agglutination, and has a role in stimulation of dendritic cells. *Journal of Medical Microbiology* **62:** 1135-1143.

Reed K, Morris T, Cronan J (1994). Mutants of Escherichia coli K12 that are resistant to a selenium analog of lipoic acid identity unknown genes in lipoate metabolism. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 3720-3724.

Richardson DJ (2000). Bacterial respiration: a flexible process for a changing environment. *Microbiology* **146:** 551-571.

Rinke C, Schwientek P, Sczyrba A, Ivanova N, Anderson I, Cheng J, Darling, A, Malfatti, S, Swan, BK, Gies, EA, Dodsworth, JA, Hedlund, BP, Tsiamis, G, Sievert, SM, Liu, WT, Eisen, JA, Hallam, SJ, Kyrpides, NC, Stepanauskas, R, Rubin, EM, Hugenholtz, P, Woyke, T (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Robinson CR, Sligar SG, Michael L. Johnson GKA (1995). Hydrostatic and osmotic pressure as tools to study macromolecular recognition. *Methods in Enzymology*. Academic Press. pp 395-427.

Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, Yooseph S Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, Yooseph S, Wu DY,

Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter, JC (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Biology* **5:** 398-431.

Sanford R, Wagner D, Wu Q, Chee-Sanford J, Thomas S, Cruz-Garcia C, Rodriguez G, Massol-Deya A, Krishnani KK, Ritalahti KM, Nissen S, Konstantinidis KT, Loffler FE (2012). Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proceedings of the National Academy of Sciences of the United States of America* **109:** 19709-19714.

Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environmental Microbiology* **11:** 2078-2093.

Shaw A, Halpern A, Beeson K, Tran B, Venter J, Martiny J (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environmental Microbiology* **10:** 2200-2210.

Singer E, Emerson D, Webb E, Barco R, Kuenen J, Nelson W Chan, CS, Comolli, LR, Ferriera, S, Johnson, J, Heidelberg, JF, Edwards, KJ (2011). Mariprofundus ferrooxydans PV-1 the First Genome of a Marine Fe(II) Oxidizing Zetaproteobacterium. *Plos One* **6**.

Smedile F, Messina E, La Cono V, Tsoy O, Monticelli L, Borghini M, Giuliano, L, Golyshin, PN, Mushegian, A, Yakimov, MM (2013). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environmental Microbiology* **15:** 167-182.

Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland E, Dashiell P, Gomez ML,
Sieracki ME, DeLong EF, Herndl G, Stepanauskas R (2011). Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. *Science* **333:** 1296-1300.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences* **110:** 11463-11468.

Thrash JC, Temperton B, Swan BK, Landry ZC, Woyke T, DeLong EF, Stepanauskas R, Giovannoni SJ (2014). Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *ISME J*.

Teira E, van Aken H, Veth C, Herndl G (2006). Archaeal uptake of enantiomeric amino acids in the meso- and bathypelagic waters of the North Atlantic. *Limnology and Oceanography* **51:** 60-69.

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, Giovannoni SJ (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452:** 741-744.

Usui K, Hiraki T, Kawamoto J, Kurihara T, Nogi Y, Kato C, Abe, F (2012). Eicosapentaenoic acid plays a role in stabilizing dynamic membrane structure in the deep-sea piezophile Shewanella violacea: A study employing high-pressure time-resolved fluorescence anisotropy measurement. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1818:** 574-583.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith, Hamilton O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, Kilpert F, Worden AZ, Giovannoni, SJ (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* **7:** 1322-1332.

Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro F, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi C, Bartlett DH, Valle G (2005). Life at depth: Photobacterium profundum genome sequence and expression analysis. *Science* **307:** 1459-1461.

Vignais P, Billoud B (2007). Occurrence, classification, and biological function of hydrogenases: An overview. *Chemical Reviews* **107:** 4206-4272.

Viklund J, Martijn J, Ettema TJG, Andersson SGE (2013). Comparative and Phylogenomic Evidence That the Alphaproteobacterium HIMB59 Is Not a Member of the Oceanic SAR11 Clade. *PLoS ONE* **8:** e78858.

Wang C, Le SY, Ali N, Siddiqui A (1995). An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5' noncoding region. *RNA* **1:** 526-537.

White D (2009). Modular Design of Li-Ion and Li-Polymer Batteries for Undersea Environments. *Marine Technology Society Journal* **43:** 115-122.

Wilson ST, del Valle DA, Segura-Noguera M, Karl DM (2014). A role for nitrite in the production of nitrous oxide in the lower euphotic zone of the oligotrophic North Pacific Ocean. *Deep Sea Research Part I: Oceanographic Research Papers* **85:** 47-55.

Yancey PH, Gerringer ME, Drazen JC, Rowden AA, Jamieson A (2014). Marine fish may be biochemically constrained from inhabiting the deepest ocean depths. *Proceedings of the National Academy of Sciences* **111:** 4461-4465.

Chapter 3

Expansion of the metabolic potential of candidate phylum OD1 based on cells obtained from the Challenger Deep, Mariana Trench

ABSTRACT

Candidate phylum OD1 is a group of uncultivated microbes characterized by reduced genomes with limited metabolic potential. In this study we analyzed thirteen single amplified genomes (SAGs) from surficial sediment samples colleted within the Challenger Deep of to the Mariana Trench at a water column depth of 10,908 m. Comparative genomics was used to examine the metabolic potential harbored by these SAGs (OD1-DSC). The OD1-DSC genomes contain additional features not previously identified in this division. This includes the presence of genes involved in lipopolysaccaride biosynthesis, components of the electron transport chain including NADH-dehydrogenase and cytochrome c oxidase, and the presence of nitrate reductase and additional genes associated with denitrification. Horizontally transferred genes were abundant, especially those associated with archaea. The results indicate that some OD1 cells are capable of much greater metabolic versatility and genetic exchange than previously ascribed to this candidate phylum.

BACKGROUND

Microbial abundance in ocean surficial sediments (0-10cm) has been estimated to be around 13% of all the microbial biomass in the ocean (Whitman *et al*. 1998). Ocean sediments harbor not only high biomass but great diversity due in part to the varied environmental conditions found throughout, e.g. surficial accumulation of organic matter and oxygen consumption, and development of stratified redox gradients (Torsvik *et al*, 2002; Zinger *et al*, 2011; Edwards *et al,* 2012). Surveys of microbial diversity in ocean sediments have included the presence of numerous candidate phyla (CP, Schauer *et al*, 2009; Nunoura *et al*, 2012).

Among these CP the group OD1 stands out as one of the most studied due to its abundance in many different anoxic marine and terrestrial environments (Elshahed *et al*, 2005; Gihring *et al*, 2011; Peura *et al*, 2012). The OD1 CP was originally described as part of the OP11 group but was later placed into its own division based on its highly divergent 16S rRNA gene sequences (Harris *et al*, 2004). Limited metabolic information about the OD1 CP has been acquired from metagenomic composite genomes or by using single cell genomics (Rinke *et al*, 2013; Wrighton *et al*, 2012; Wrighton *et al*, 2014).

Rinke and colleagues proposed the superphylum Patescibacteria that encompasses the phyla OD1 (Parcubacteri), Microgenomates (OP11) and Gracilibacteri (GN02). The name Patescibateria reflects their reduced metabolic potential. *Candidatus Paceibacter normanii* (single cell AAA255-P19) was designated as a *Candidatus* type species for the OD1 CP (Rinke *et al*, 2013). It was recovered from brackish water present at 120 m depth in Sakinaw Lake, British Columbia, Canada. Its genome is 0.6Mbp in size and estimated to be 70% complete (Rinke *et al*, 2013). *Candidatus Paceibacter normanii*

appears to have very limited metabolic potential highlighted by the lack of genes involved in sugar and amino acid degradation, the pentose phosphate pathway, pyruvate metabolism or the electron transport chain.

In 2012, Wrighton and colleagues published a general description of 21 OD1 genomes recovered from a metagenome of an acetate-amended aquifer, along with another 28 genomes from other novel CP (Wrighton *et al*, 2012). These findings included the report of a new sublineage, OD1-i which was discovered to possess a relatively reduced genome in terms of metabolic potential, lacking the tricarboxylic acid cycle (TCA) cycle and oxidative phosphorylation components. With a mostly fermentative metabolism it was predicted to utilize acetyl-CoA synthetase for ATP generation and to reoxidize NADH produced during glycolysis by converting pyruvate to D-lactate and acetyl-CoA to ethanol.

More recently, metagenomic analyses have been conducted on a microbial community present in a sediment column biostimulated with acetate-amended ground water. Acetate stimulation resulted in a succession of species that changed with the availability of consumed and generated nutrients. OD1 had the highest relative abundance before sulfate reduction occurred (Kantor *et al*, 2013). A nearly complete OD1 genome sequence, was recovered via genome reconstruction from the metagenome, and its microbe source was designated RAAC4. This genome sequence information has reinforced the conclusion that members of the OD1 group have limited metabolic potential. RAAC4 lacks a TCA cycle and respiratory chain enzymes and appears to be a strictly fermentative anaerobic organism. It does not contain genes involved in the conversion of pyruvate to acetyl-CoA or for the utilization of acetyl-CoA, and it lacks

biosynthetic genes for nucleotides, lipids and most amino acids. However, it does contain genes associated with the pentose phosphate pathway and a modified Embden-Meyerhof-Parnas (EMP), and it does appear to be able to utilize complex organic carbon and perhaps to create biofilms.

During the Deepsea Challenge Expedition a push core as obtained by the manned submersible Deepsea Challenger within the Challenger Deep, the deepest ocean location on earth. Located in the western Pacific Ocean, it extends to a depth of approximately 10,920 m (Nakanishi and Hashimoto, 2011), corresponding to about 110 megapascals ([MPa], 1,090 atmospheres, and 16,000 pounds per square inch) of hydrostatic pressure. Single cell-derived genomes were obtained from the Challenger Deep pushcore sample and among the genomes identified were thirteen associated with the OD1 CP. These genomes provided the opportunity to examine the evolution and adaptation of the OD1 CP in the content of an extreme habitat.

Elucidating the metabolic capabilities of novel microorganisms, especially those belonging to a candidate division, is clearly of importance to the understanding of the biogeochemical cycling of carbon and other nutrients in the ultradeep ocean. Here we present a comparative genomic analysis of the 13 single cell genomes and one combined genome assembly. The results indicate that members of the OD1 CP from the Challenger Deep have greater metabolic potential than previously reported for this CP, including the presence of genes involved in lipopolysaccharide biosynthesis, oxidative phosphorylation, and nitrate reduction.

MATERIALS AND METHODS

Collection and sorting

Sediments were collected at from a depth of 10,908 m using a push-core apparatus controlled by a hydraulic arm within the manned submersible Deepsea Challenger Sampling occurred on March 26, 2012 in the "East Deep" (Fujioka *et al*, 2002) of the Challenger Deep at 142.59° E, 11.37° N during the Deepsea Challenge Expedition. Recovered sediment was placed in glycerol/TE buffer (Rinke *et al*, 2014) and first stored in liquid nitrogen at -196ºC and later in an ultralow freezer at –80ºC prior to single cell sorting. Samples were transferred to the J. Craig Venter Institute (JCVI) for sorting. The sediment sample was gently vortexed and allowed to settle briefly before filtering through a 35μm mesh (BD Biosciences, San Jose, CA, USA) to avoid larger sediment particles. Cells were stained with 10x SYBR Green I nucleic acid stain (Invitrogen, Carlsbad, CA, USA). Single cells were sorted using a cooled FACS-Aria II flow cytometer (BD Biosciences, San Jose, CA) and microtiter plates were stored at –80ºC until further processed.

Genome amplification and sequencing

DNA was amplified using a custom BioCel robotic system (Agilent Technologies, Santa Clara, CA) as described by McLean *et al* (2013). Genomic material in the sorted microbial cells was amplified by multiple displacement amplification (MDA) in a 384-well format using a GenomiPhi kit (GE Healthcare, Waukesha, WI, USA). 16S rRNA genes were PCR amplified from diluted MDA products using universal bacterial primers 27F and 1492R (Weisburg *et al*, 1991) as follows: 94 °C for 3 min, 35 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 90 s, and 72 °C for 10 min. PCR products were treated

with exonuclease I and shrimp alkaline phosphatase (Thermo Fisher Scientific Inc., Waltham, MA, USA) and 16S rRNA gene amplicons were sent for Sanger sequencing at the Joint Technology Center (JTC, J. Craig Venter Institute, Rockville, MD, USA). 16S rRNA gene trace files were analyzed and trimmed with the CLC Workbench software program (CLC Bio, Cambridge, MA, USA). Chromatogram quality was assessed manually, and PCR product sequences with both forward and reverse sequencing primer reads of poor quality were excluded from further analysis. Resulting 16S rRNA gene sequences were evaluated for evidence of microbial DNA contamination associated to MDA reagents, and any samples judged to be contaminated were removed from consideration for whole genome sequencing. These curated sequences were then compared to the NCBI nr/nt database using BLASTN (Altschul *et al*, 1990) in order to perform initial taxonomic characterization of the sorted cells. DNA recovered from 76 cells was prepared for Illumina sequencing. Libraries were prepared using the multiple barcode technology of the Nextera™ DNA Sample Prep Kit (Illumina, San Diego, CA, USA) and sent to JTC for sequencing. After sequencing, samples were de-multiplexed to separate barcoded sequences for each corresponding single cell genome.

Assembly, annotation and genome completion

Sequences were assembled using the Spades assembler, SPAdes 3 (Bankevich *et al*, 2012). Genomes were processed using Nesoni (www.vicbioinformatics.com/software.nesoni.shtml) and annotated by IMG-ER (https://img.jgi.doe.gov/cgi-bin/er/main.cgi, Markowitz *et al*, 2014) for complete genome annotation.

16S rRNA gene sequences recovered from each SAG were analyzed by BLASTn against the NCBI nr/nt database (Altschul *et al*, 1990). Sequences with 85% or greater similarity to OD1 CP were extracted and used for phylogenetic reconstruction, along with sequences previously described belonging to the OD1 CP from previous publications (Wrighton *et al*, 2012; Kantor *et al*, 2013; Rinke *et al*, 2013). All sequences extracted from NCBI were also annotated with regard to their associated environmental source, and when it was available, seawater depth. Sequences were aligned with the SINA aligner (http://www.arb-silva.de/aligner/, Pruesse *et al*, 2012) and maximum-likelihood tree were created using FastTree (Price *et al*, 2009).

Genome-encoded protein predictions were obtained from IMG-ER and classified phylogenomically using DarkHorse software, version 1.4 (http://darkhorse.ucsd.edu/, Podell and Gaasterland, 2007). DarkHorse was used to predict horizontally transferred genes by assigning a probability that a given encoded protein belonged to the genome being investigated. DarkHorse results were also used to identify potential contaminating sequences among SAG contigs, based on whether or not taxonomic lineages associated with predicted proteins on each assembled contig were similar to or different from the rest of the contigs (Jones *et al,* 2011). Estimated genome completeness for each SAG was calculated as previously described by Rinke *et al* (2013) by using universal single-copy genes. Functional comparisons were performed using the IMG-ER platform (Markowitz *et al*, 2014).

Results and Discussion

Genomic properties

From the sediment samples 3520 total cells were sorted, 704 cells were subjected to MDA, and 494 MDA reactions were positively amplified as identified by subsequent 16S rRNA gene polymerase chain reaction amplification. OD1 genomes represented approximately 5.4% of the totally phylogenetic community within this sample based on the fraction of singe cell- derived 16S rRNA gene sequences ascribed to this CP (Supplementary Figure 3.4). Thirteen genomes that belong to the OD1 CP were analyzed (Table 3.1). Sequences recovered ranged from 0.3 to 1.1 Mbp and genome completeness ranged from 28% (0.5Mbp) to 88% (0.8 Mbp). Open reading frame predictions ranged from 399 to 1365 genes and predicted proteins ranged from 218 and 795 per genome. Percent GC ranged from 35% to 45%, where the most abundant percentage was 38% GC, and the percent of coding sequence ranged from 75 to 88%. 16S rRNA genes were recovered from all but two genomes and tRNA counts ranges from 15 to 40 per genome. The number of conserved hypothetical proteins ranged from 7 (2% of the genome) to 38 (5% of the genome). One combined assembly was conducted by combining amplified DNA from two of the single cells (OD1_DSC11 and OD1_DSC12), which were 100% similar at average nucleotide identity (ANI) level. It resulted in 75% genome completeness and 1434 predicted genes.

Phylogenetic relationships

For the twelve genomes for which 16S rRNA genes were recovered in the sequenced genome, phylogenetic relationships place them within three major clades within the OD1 CP (Figure 3.1). Figure 3.1 presents a phylogenetic tree of the OD1 CP

including the OD1-DSC SAGs. Orange lines denote samples from this study, purple lines represent samples previously described and reference sequences from amended subterranean aquifers study (Wrighton *et al*, 2012) and green lines represent other single cell genomes from brackish waters (Rinke *et al,* 2013). Environmental sequences with at least 85% 16S rRNA gene similarity to OD1-DSC SAGs were incorporated into the phylogenetic tree analysis (Figure 3.1). Their phylogenetic distribution suggests that the OD1-DSC are divided into three major clades. Following the nomenclature convention of Wrighton *et al* for OD1-i (Wrighton *et al*, 2012), we have designated these additional clades OD1-ii, OD1-iii, and OD1-iv. Clade OD1-ii is potentially a large lineage of microbes and encompasses most of the previously reported OD1 sequences, divided among various subclades, and includes *Candidatus Paceibacter normanii* and RAAC4. Clade OD1-iv does not possess previously described OD1 sequences and Clade OD1-iii which is divided into two subclades, with previously described sequences falling within one of the clades and some of the OD1-DSC genomes, along with some environmental sequences, falling within the other subclade. OD1-DSC9 and 11 are the only single cells most similar to *Candidatus Paceibacter normanii* (Rinke *et al,* 2013). None of OD1-DSC cells cluster directly with organisms that are closely related to RAAC4. All of the other OD1 cells do not cluster directly with previously characterized OD1 microorganisms. Clade OD1-iii and OD1-iv appear to represent a new lineage within the CP based on 16S rRNA phylogeny, and within clade OD1-ii novel subgroups lacking previously characterized OD1 microorganisms are also present. Phylogenetic associations were further investigated by analyzing other conserved phylogenetic gene markers and the results mostly confirmed the 16S rRNA base relationship although the distinction

between clade OD1-iii and OD1-iv cannot always be reproduced (see Supplementary Figures S3.5, S3.6 and S3.7). This may in part be due to missing marker genes in some of the clade OD1-iii and OD1-iv genomes.

When comparing the 16S rRNA gene to the NCBI nr database using BLASTN the top hit percent identity for each OD1-DSC SAG 16S rRNA gene ranged from 83 – 97% similarity to environmental sequences. The only 97% identity score was for a deep-sea subsurface sample at a water depth level >5000m. Average nucleotide identity (ANI) relationships among the DSC SAGs ranged from 85 –100% similarity at the whole genome level, and 92 – 100% at the coding sequences level.

Metabolic profiles

All OD1-DSC SAG genomes were assembled and annotated separately but analyzed together, with a focus on their metabolic pathways (Figure 3.3; Table S3.4). KEGG pathway metabolic reconstruction of the DSC OD1 CP genomes suggests that DSC OD1 cells are mostly heterotrophic microbes with limited chemoheterotrophic substrate utilization ability along with capabilities for aerobic, anaerobic and fermentative metabolism. As a community the DSC genomes from clades OD1-ii, OD1-iii, and OD1-iv have the potential for the Embden–Meyerhof–Parnas (EMP) pathway, which is the main pathway for the conversion of glucose to pyruvate in order to generate energy. All EMP genes but one are present, the exception being phosphofructokinase (PFK), which is not found in any of the DSC genomes. A partial or complete EMP pathway is present in other OD1 genomes. OD1-i and AAA255-P19 also have all of the main EMP enzymes with the exception of PFK, and RAAC4 has all of the EMP enzymes including PFK. The

lack of PFK in most OD1 genomes could reflect a primary EMP role for gluconeogenesis rather than glycolysis, using this anabolic pathway to synthesize sugar molecules from pyruvate.

Genes encoding for enzymes in the tricarboxylic acid cycle (TCA) are scarce but present in the clade OD1-iii DSC single cell genomes, which has not previously been reported for other OD1 genomes. The TCA cycle is the process by which pyruvate is converted to carbon dioxide (CO_2) to generate energy. The cycle enzymes identified include the following: 2-oxoglutarate: ferredoxin oxidoreductase (KorA and KorB; OD1-DSC5), succinyl-CoA synthetase (OD1-DSC6 and DSC7) and citrate synthase (OD1-DSC6). Many other enzymes that are required for a complete TCA cycle are not present in any of the OD1 DSC genomes, including those affiliated with OD1-iii. Although all of the DSC single cell genomes are incomplete, the most parsimonious explanation is that none of the DSC OD1 genomes encode for an entire TCA cycle.

All enzymes involved in the pentose phosphate pathways (PP), except for 6-phosphogluconolactonase, were found in the OD1-DSC genomes from clade OD1-ii, as is the case with some of the previously described OD1 genomes. The PP pathway is generates NADPH and pentose, which are primarily utilized for anabolic purposes. One enzyme, Transketolase (EC: 2.2.1.1), was found DSC7, which belongs to the OD1-iii clade. No PP related enzymes were found from clade OD1-iv. This may be due to the fact that the one genome present in this clade is only 50% complete.

The DSC OD1 genomes share with AAA255-P19, but not OD1-i and RAAC4, the presence of genes involved in the biosynthesis of purines and pyrimidines (Table S3.4). Fermentative metabolism is also present in the DSC genomes within clades OD1-ii

and OD1-iii using lactate and alcohol dehydrogenases (OD1-DSC1, 4, 6, 7, 8, 9, 11 and13).

Within clade OD1-iii, DSC6 encodes for a complete nitrate reductase operon, while the DSC2 genome within clade OD1-iv encodes for subunits of a nitrite reductase. In addition, a number of nitrate/nitrite transporters are present in clade OD1-iii (DSC5 and DSC6). This suggests that some member of clade OD1-iii and perhaps OD1-iv may be able to respire nitrate and/or nitrite as a terminal electron acceptor, counter to previously suggested descriptions that all OD1 may be strictly fermentative-anaerobes (Wrighton *et al,* 2012). The nitrate reductase operon is most closely related to an operon found in members of the *Moraxella* genus within the gammaproteobacteria. The OD1-DSC6 also genome encodes other proteins and enzymes, like NADH dehydrogenase (first step in the respiratory chain for oxygen an nitrate respiration), cytochrome c oxidase (oxygen-reducing terminal oxidase), a nitrous oxide reductase (reduces of nitrous oxide into nitrogen) and V-ATPase (possible ATP synthase), that suggest that this genome is potentially involved complex respiration of both nitrate and oxygen (Chen and Strous, 2012).

Another DSC OD1 feature not previously identified within members of this CP is the presence of genes associated with adaptation to oxidative stress. These features span DSC OD1-ii, OD1-iii and OD1-iv and include peroxiredoxin (DSC1, 3, 4, 6, 7 11 and 12), and catalase (DSC4 and 7). The presence of these enzymes indicates some degree of oxygen tolerance, including the possibility that DSC OD1 cells are not obligate anaerobes but rather facultative anaerobes.

In contrast with the notion that cells within the OD1 phylum lack the machinery

for the electron transport chain (Wrighton *et al,* 2012), DSC genomes possess many of the components associated with microbial respiration by oxidative phosphorylation. Among the genomes from clade OD1-iii are genes coding for heme/copper-type cytochrome/quinol oxidases (cytochrome C oxidase; DSC6 and 7), which is involved in oxidative phosphorylation, utilizing oxygen as a terminal electron acceptor. F type ATPases, whose main role is to catalyze the synthesis of ATP using energy generated by cellular respiration (Yoshida *et al*, 2001), are found in members of clades OD1-ii and OD1-iii (DSC1, 4, 5, 7, 11 and 12). A V type ATPase, most commonly known as a strictly ATP hydrolysis enzyme but in some microorganisms is also able to synthesize, ATP (Toei *et al*, 2007) is encoded in one genome within the clade OD1-iii (DSC6). NADH-dehydrogenase (ubiquinone), the enzyme responsible for the first step in the electron transport chain, is also found in clades OD1-ii and OD1-iii (DSC1, 4, 6, 11, 12 and 13). Cytochrome c1, part of complex III of the electron transport chain, is only found in a genome from clade OD1-iii (DSC7) (Figure 3.3, Table S3.4).

The DSC OD1 genomes also reveal new information about OD1 CP cell surface structure. As expected, most of the DSC OD1genomes contain genes involved in peptidoglycan biosynthesis, but in contrast to previous findings, the cell architecture of the DSC OD1 cells appear to also include lipopolysaccharide, a defining feature of Gram negative bacteria (Figure 3.3). Genomes from the DSC clades OD1-ii, OD1-iii and OD1-iv all have the potential for lipopolysaccaride elaboration including the presence of genes for lipid A core-O-antigen ligase and related enzymes, 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase; EC:2.7.7.38; *kdsB*), 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase (KDO 8-P phosphatase; EC:3.1.3.45; *kdsC*),

glycosyltransferases (*pimB*), as well as transport proteins for o-antigen and o-antigen ligase. This leads to the conclusion that members of the OD1-DSC have gram-negative cell-like cell membranes.

Another notable surface property is the presence of pili associated with surface movement. Genes used for twitching motility and type IV pili biosynthesis are present in genomes from clades OD1-ii and OD1-iii (OD1-DSC1, 3, 4, 5, 6, 7, 9, 10, 11 and 13). Type IV pili biosynthesis has previously been described for OD1-i, RAAC4 and AAA255-P19, and it is one of the two characteristic shared between all the OD1-DSC genomes. The prevalence of type IV pili biosynthesis was also observed in all of the partial OD1 genomes examined by Rinke and colleagues (Rinke *et al*, 2013). The Type IV pilus system is a multifunctional machine used for adherence, motility, DNA transfer, protein secretion and can even act as a nanowire carrying electric current (Shi and Sun, 2002, Melville and Craig, 2013). Genomes from all clades also encode genes involved in the type II secretory pathway, which also have high homology to type IV pilus biosynthetic genes (Ayers *et al*, 2010). The type II secretory pathway is responsible for the secretion of hydrolytic enzymes and it is different from other secretion systems because it mainly secretes folded proteins (Sandkvist, 2001, Korotkov *et al,* 2012). For example enzymes involve in cellulose degradation, like beta-glucosidase, have been reported to be secreted by type II secretion system (Gardner and Keating, 2010).

Complex carbon degradation is another characteristic that the DSC OD1-ii and OD1-iii genomes share with other described OD1 genomes. This includes a number of glycosyl hydrolases (DSC1, 3, 6, 11 and 13) and a b-glucosidase involved in the degradation of cellulose. (DSC4, 7 11, 12 and 13). Glycosyl hydrolases (RAAC4) and a

b-glucosidase gene, although less abundant, are also found in RAAC4 and AAA255-P19, respectively. The ability to produce enzymes that can degrade recalcitrant organic matter could serve as an adaptation to deep-sea environments (Nagata *et al*, 2010), including the ultradeep setting of the Challenger Deep (Kobayashi *et al*, 2012; Lauro *et al*, 2013; Glud *et al*, 2013).

Environmental sensing and regulation

The DSC OD1 cells encode both heat shock and cold shock proteins. Some of these proteins are also encoded by the genomes RAAC4 and AAA255-P19 as is the case of the heat shock proteins DnaK/DnaJ and GrpE, but cold shock proteins seem to be less prevalent in RAAC4 and AAA255-P19. The heat shock system is an early-evolved system of protein-folding proteins, both molecular chaperones and protein-folding catalysts, linked to transcription factors that generate cellular responses to external and internal protein unfolding stresses (Feder and Hofmann, 1999). The heat shock system functions encoded within the DSC genomes from all three DSC clades are DnaK/DnaJ, GrpE, GroEL/GroES, small heat shock protein and ClpB (Lindquist, 1986; Richter *et al*, 2010). Cold shock proteins (CSP) such as CspA, are found in the SAGs from all DSC clades (DSC1, 2, 3, 4, 6, 7, 9, 11, 12 and 13). These proteins facilitate adaptation to low temperatures (Ivancic *et al*, 2013). For example, CspA prevents the formation of inhibitory acts as an RNA chaperone to keep mRNAs free of secondary structure (Phadtare and Inouye, 1999).

Sigma factors (σ) are an essential part of the transcription equation. For RNA polymerase to begin transcription at a particular promoter, it must first interact with a σ

subunit to form an active RNA polymerase holoenzyme. The σ subunit has three main functions: to ensure the recognition of specific promoter sequences; to position the RNA polymerase holoenzyme at a target promoter; and to facilitate unwinding of the DNA duplex near the transcript start site (Browning *et al,* 2004). A number of regulatory sigma factors also exist in the DSC genomes. Not surprisingly this includes the essential sigma factor, σ^{70}, which is responsible for the transcription of most genes expressed in exponentially growing cells (Wösten, 1998). Among the non-essential sigma factor found in the OD1-DSC genomes from all three clades are two involved in heat shock response, σ^{E} (σ^{24}; DSC2, -DSC3, DSC6, DSC10, DSC12) and σ^{32} (DSC5) (Yura *et al*, 1993; Raina *et al*, 1995; Alba and Gross, 2004; Wade *et al*, 2006). The σ^{E} regulatory system is used for high pressure and cold temperature adaptation in the deep-sea bacterium *Photobacterium profundum* (Chi and Bartlett, 2004). Curiously, σ^{E} appears to be more widespread than σ^{32} in the DSC genomes, perhaps because of its role in adaptation to deep-sea stressors.

σ^{54} is found in the DSC6 genome from clade OD1-iii, which directs the transcription of genes involved in a variety of physiological processes responsive to nutrient limitation, including nitrogen assimilation and fixation, substrate-specific transport systems, and utilization of alternative carbon and energy sources (Merrick, 1993).

The ability of the DSC cells to cope with environmental stress is also reflected in their proteases. Among the different kinds of proteases present in the SAGs is endoprotease ATP-dependent Clp protease associated with peptide degradation under

heat shock conditions (Porankiewicz *et al,* 1998), ATP-dependent metalloprotease FtsH associated with the degradation of heat shock sigma factor σ^{32} (Tomoyasu *et al*, 1995) and PrsW family protease associated with the degradation of anti-sigma factors that control the function of extracytoplasmic function (ECF) σ factors which in turn are associated with cell membrane stress (Ho and Ellermeier, 2011).

Horizontally transferred genes

The OD1 group is proposed to be a large and potentially diverse CP. To assess the diversity across the DSC SAGs, the predicted proteins encoded within each was compared to the NCBI database by BLAST. The top hit predictions were extracted and classified based on their taxonomic association at the genus level. The most abundant top hit for all of the genomes was *Candidatus Paceibacter normanii (*Rinke *et al*, 2013), reflecting a common genome core across the entire phylum. Some of the most abundant non-OD1 top BLAST hits are to various species of *Clostridium and Bacillus* (gram-positive, spore-forming microorganisms within the phylum *Firmicutes*), *Candidatus Saccharimonas aalborgensis* (anaerobic, gram-positive, sugar fermenting microorganisms within the TM7 or Saccharibacteria phylum; Albertsen *et al,* 2013), and to *Dehalococcoides mccartyi (*anaerobic microorganism with an obligate requirement for reductive dehalogenation; Löffler *et al*, 2013). The fact that the most abundant non *Candidatus Paceibacter normanii* matches for each genome appear to be shared among most of the SAGs suggests that they have been vertically acquired. However, differences in gene content do exist among the DSC SAGs. This is shown by the non-metric multidimensional scaling (nMDS) plot that presents the overall similarity of gene content

populations among each of the DSC SAGs (Figure 3.2). Although this evaluation of the total genome similarity, encompassing both vertically and laterally transferred genes, is limited by the incompleteness of the genome sequences, the results are consistent with the proposed phylogenetic associations between the SAGs. Extensive variation in genes introduced by horizontal gene transfer among the genes are also represented. For example, while the OD1-iii and OD1-iv clades (Figure 3.2, blue) are separated from the OD1-ii members (Figure 3.2, orange and red), a cluster OD1-ii genomes, namely DSC8, 9 and 11 are removed from its other OD1-ii subclade members. This separation resembles in a way their phylogenetic association as they appear to be closely related by 16S rRNA phylogeny, but at the same time their genomic composition separates them from the rest of the OD1-ii clade. DSC8, for which it was not possible to determined phylogeny based on 16S rRNA, falls in the same cluster are DSC9 and 11, suggesting that DSC8 may also be a member of the OD1-ii clade.

All OD1-DSC genomes encode genes that appear to have been horizontally transferred from archaea and eukarya (Table 3.2). This is based on the lineage probability index measurements for each gene using the DarkHorse program (Podell and Gaasterland, 2007). The LPI index evaluates the probability of genes encoding proteins to have been horizontally transferred. The larger the LPI the less likely it is for the gene coding for the predicted protein to have been horizontally transferred. In the case of horizontally transferred genes from bacteria, the LPI values were too high (> 0.6) to discern HGT. This may be in part due to the lack of available sequenced genomes from the OD1 CP. In total, within all 13 genomes, 42 predicted proteins are more closely related to sequences present in members of the eukarya and 329 to members of the

archaea. Of the predicted proteins most closely related to those encoded within archaea 78% are associated with the phylum *Euryarchaeota*, while 40% of the genes whose protein sequences appear to have been transferred from eukarya are most related to those present in *Metazoa*. Given that many of the putative HTGs are shared among OD1 single cells belonging to separate phylogenetic subgroups, it is likely that they encode a selective advantage. Sixty five percent have a predicted function. Among the most abundant horizontally transferred genes (HTG) are glycosyltransferases, which are responsible for glycosidic bond formation in glycoconjugates such as polysaccharides, lipopolysaccharides, peptidoglycan, glycoproteins, etc. (DSC1, 3, 4, 5, 6, 7, 9,10, 12 and 13; Lairson *et al*, 2008). Other genes are associated with additional aspects of lipopolysaccharide biosynthesis, oxidative stress adaptation and gluconeogenesis.

One example of a HGT shared among two or more DSC genomes is the gene for ADP-ribose pyrophosphatase. ADP-ribose pyrophosphatase is part of the NUDIX hydrolase family of enzymes that perform the catalysis of ADP-ribose to AMP and ribose-5-P, which is associated with cellular maintenance, detoxification of nonenzymatic ADP ribosylation products, and tellurite resistance (Bessman *et al*, 1996; Dunn *et al*, 1999; Gabelli *et al*, 2001). These genes appear to have been horizontally transferred to six of the genomes. In some cases it is most closely related to archaea (DSC2, 3, 5 and 8) and in the case of two DSC genomes appears to have come from members of the eukarya (DSC4 and10).

Many of the genes that appear to have originated in archaea are associated with respiration. This includes a gene encoding cytochrome c biogenesis protein, which required for cytochrome C biosynthesis (DSC1, 7, 12 and 13), heme/copper-type

cytochrome/quinol oxidases subunit 2, which is a subunit within the cytochrome C oxidase (DSC5 and 6) a methylase involved in ubiquinone/menaquinone biosynthesis that catalyzes the methylation of the ubiquinone (coenzyme Q) and menaquinone (vitamin K2), components of the respiratory chain (DSC6, and 10) (Lee *et al*, 1997), and plastocyanin (OD1-DSC1, 3, 10, 12 and 13). Platocyanins are blue copper proteins that have been mostly studied for their role in photosynthesis in cyanobacteria, algae and plants (Redinbo *et al*, 1993). However, plastocyanin-like proteins are also present in archaea, including the ammonia oxidizing archaeon *Nitrosopumilus maritimus*, where it is thought to shuttle electrons among the complex II components of the electron transport chain, including a cytochrome-c-like protein (Walker *et al*, 2010). Additional archaeal – linked genes that may be indirectly linked to respiration are the peroxiredoxin and alkyl hydroperoxide reductase enzymes involved in the removal of peroxide/oxygen radicals resulting from oxygen respiration (DSC1, 6, 7, 8, 9 and 11),

Among the HTGs that appear to have eukarya donors are those associated with DNA and tRNA synthesis, cell detoxification and fatty acid metabolism. Some OD1-DSC also have annotated phage-like sequences (Table 3.3) many of them associated with phage coat protein or replication proteins, but interestingly the type II secretory pathway, component PulD, found in 3 of the OD1-DSC genomes (DSC2, 5 and 12) is most closely related to enterobacterial phage M13 from the Inoviridae family (Table 3.3).

OD1 cells are a significant fractions of the microbial population present in the surficial sediments of the Challenger Deep, as is the situation in many other suboxic and anoxic environments. The results of this study reinforce the view of the OD1 CP as organisms with small genomes that are able to metabolize organics by fermentation.

However, it also expands the current knowledge of the metabolic potential associated with the OD1 CP. The novel genes discovered, many of archaeal or eukaryal origin, illustrate for the first time the potential for lipopolysaccharide biosynthesis, aerobic and anaerobic respiration, the possibility for the utilization of complex organics and the presence ofsensory/response systems for coping with environmental changes. These new functions suggest that these microorganisms are much more metabolically active and environmentally responsive than previously indicated, warranting a reassessment of the degree of reduction in metabolic potential present in this CP. It remains to be determined how many of the new functions identified are unique to the selective pressure of deep ocean environments.

Acknowledgements

We are grateful for the financial support provided by the National Science Foundation (0801973 and 0827051), a National Science Foundation Graduate Research Fellowship (068775), the National Aeronautics and Space Administration (NNX11AG10G), a National Institutes of Health Marine Biotechnology Training grant (T32GM067550) and a gift from Earthship Productions. We are specially grateful to James Cameron for his contribution to the collection of these samples.

Chapter 3 is a full-length manuscript in preparation for publication: Rosa León Zayas, Logan Peoples, Sheila Podell, Mark Novotny, Roger S. Lasken and Douglas H. Bartlett. 'Expansion of the metabolic potential of candidate phylum OD1 based on cells obtained from the Challenger Deep, Mariana Trench' with permission from all coauthors

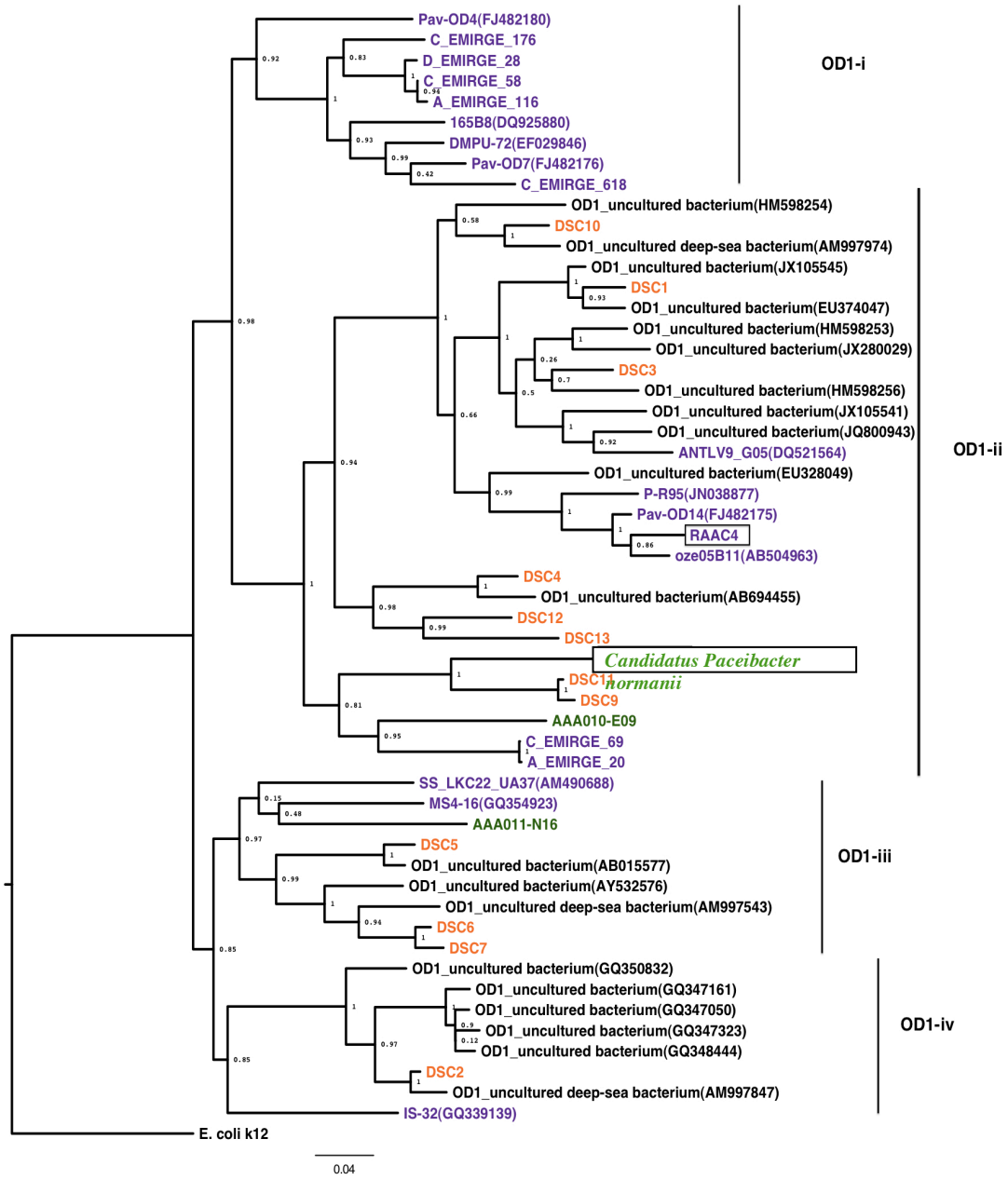

Figure 3.1 Phylogenetic tree of 16S rRNA gene from of OD1-DSC SAGs
Rooted maximum likelihood phylogenetic tree of 16S rRNA gene for eleven single amplified genome (SAGs) and related cultured and uncultured organisms are shown. OD1-DSC are highlighted with orange lines, purple lines represent samples previously described from amended subterranean aquifers (Wrighton *et al*, 2012) and green lines represent other single cell genomes from brackish waters (Rinke *et al*, 2013). Four clades are highlighted clade OD1-i, OD1-ii, OD1-iii and OD1-iv. Scale bar represents 0.04 changes per position. Confidence values are shown at the tree nodes.

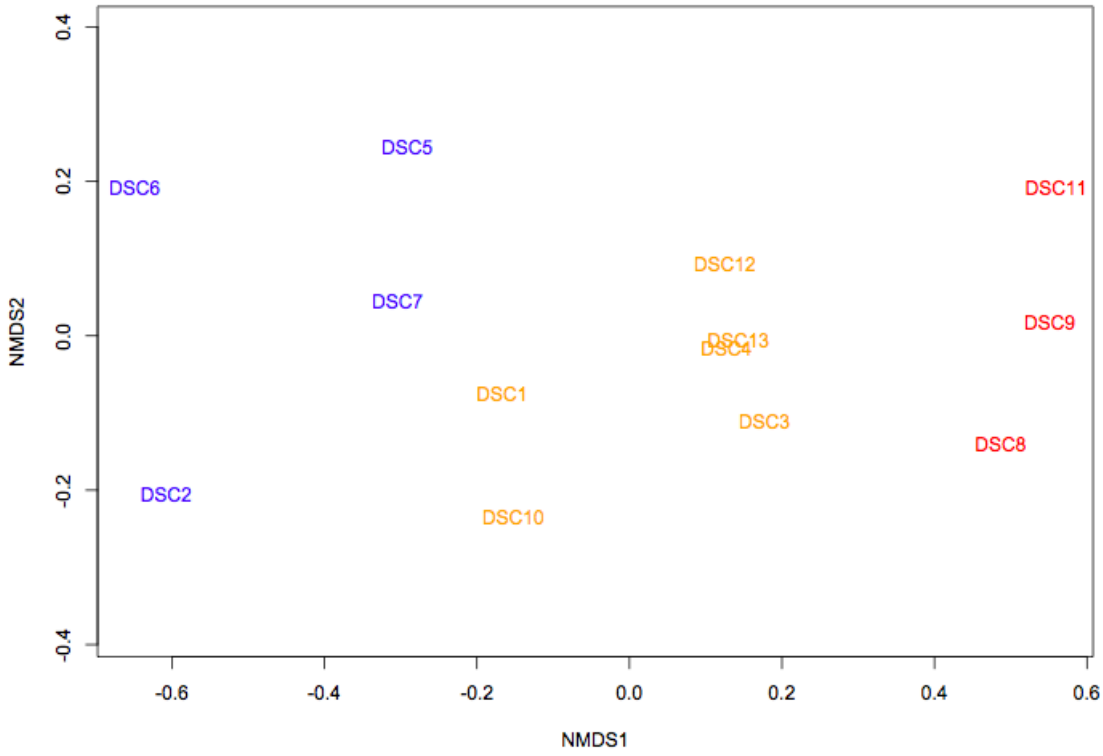

Figure 3.2 Non-Metric Multidimensional Scaling of top species hit OD1-DSC genomes
Data for the taxonomic association of each predicted protein was retrieved from the DarkHorse BLAST analysis. The abundance of each predicted protein top hit microorganism match was calculated and the most abundant organisms (44) were used to assess the similarities among the 13 DSC genomes. A Non-Metric Multidimensional Scaling (nMDS) ordination was calculated using the R-package *Vegan* a Bray-Curtis algorithm. Top hit values for each genome were normalized for the total number of proteins analyzed by DarkHorse genome. Based on their similarity matrix the data was grouped in three clusters a color-coded based on their relatedness. Three colors were used: blue, orange and red. Organisms in the blue cluster belong to clades OD1-iii and OD1-iv, while organisms in the orange and red clusters belong to clade OD1-ii. The cluters on the ordination plot mirrors that of their 16S rRNA phylogeny. Stress value is 0.049, which provided a measure of the fit of the data reported on a range of 0-1. Ordinations with stress higher than 0.3 can't be reliably interpreted; lower stress means the solution fits the data better

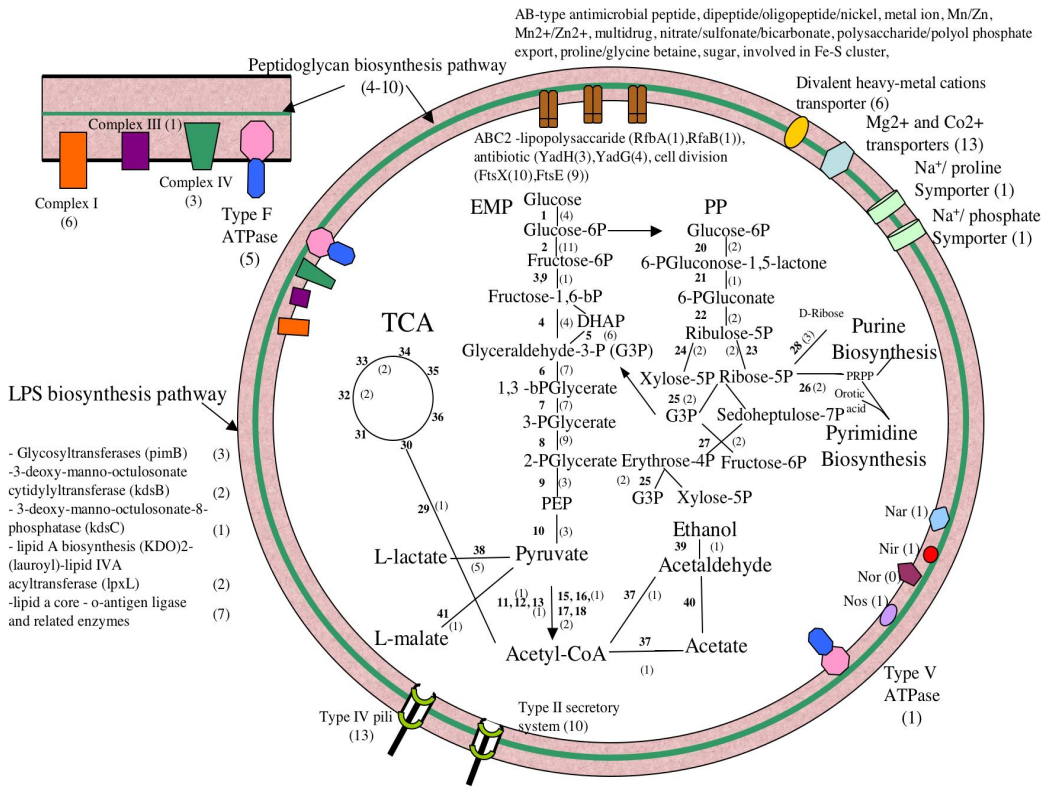

Figure 3.3 Composite metabolic potential of OD1-DSC genomes
Most enzymes highlighted in this figure are represented by bold numbers located in the lines that connect the different substrates. The number of genomes that encode for each enzyme is found in parenthesis besides the enzyme number. The enzyme number corresponds to a number found on the table supplementary Table 3.4 (Table S3.4), along with enzyme name and the genomes were the enzyme is encoded for. Other metabolic properties noted in the figure like a full list of transporter, components of the respiratory machinery and lipopolysaccaride biosynthesis can also be found in Table S3.4. EMP: Embden–Meyerhof–Parnas, PP: Phosphate Pentose, TCA: tricarboxylic acid cycle, Nar: nitrate reductase, Nir: nitrite reductase, Nor: nitric reductase, Nos: nitrous oxide reductase.

Table 3.1 Genomic properties of 13 OD1-DSC SAG genomes
Sequenced genome size, % completeness, GC%, 16s rRNA gene count, tRNA count, scaffold count, coding base % and genes count with % of function prediction and horizontally transferred genes from archaea and eukaryotic hosts are displayed for all 13 OD1-DSC genomes

Study Name	Genome Size	% Complete	GC %	16S rRNA	tRNA Count	Scaffold Count	Coding Base Count %	Gene Count	w/ Func Pred %	HTG (Euk_Arch)
OD1DSC1	884173	80	38	1	34	242	88.37	1086	58.01	40
OD1DSC2	613568	50	43	1	20	293	86.66	890	37.53	16
OD1DSC3	502208	68	38	1	35	91	88.21	630	58.73	25
OD1DSC4	655916	55	42	1	33	196	87.39	785	62.04	40
OD1DSC5	507143	26	35	1	30	395	86.23	702	50	23
OD1DSC6	1124777	77	38	1	40	352	86.35	1365	58.24	48
OD1DSC7	862086	67	38	1	22	400	86.07	1102	57.53	44
OD1DSC8	289151	23	40	0	19	166	75.17	399	54.64	18
OD1DSC9	441441	42	41	0	18	206	85.21	593	51.94	17
OD1DSC10	504962	39	38	1	15	122	86.04	592	58.45	31
OD1DSC11	578069	52	39	1	39	119	88.58	715	61.82	22
OD1DSC12	707548	65	45	1	31	366	83.28	994	56.24	33
OD1DSC13	748022	73	35	1	37	236	86.25	852	59.27	34

Table 3.2 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes

Genes with best BLAST matches to archaea or eukarya are displayed in this table including the product name annotation by IMG, the top hit species and the top hit BLAST match assessed by the DarkHorse analysis, for the genes highlighted within the article. For a complete table of horizontally transferred genes see Supplementary figure 3.3

OD1-DSC product name	Top Hit Species ARCH	Top Hit Best Blast
OD1-DSC1		
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome c biogenesis protein transmembrane region
Alkyl hydroperoxide reductase, large subunit	Thermococcus sibiricus MM 739	Glutaredoxin/thioredoxin-like protein
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
Glycosyltransferase	Methanothermococcus okinawensis IH1	group 1 glycosyl transferase
hypothetical protein	Halococcus hamelinensis	hypothetical protein
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
Plastocyanin	Candidatus Nanosalina sp. J07AB43	plastocyanin
OD1-DSC2		
ADP-ribose pyrophosphatase	Methanocaldococcus fervens AG86	NUDIX hydrolase
OD1-DSC3		
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanocella conradii HZ254	glycosyltransferase
Lipid A core - O-antigen ligase and related enzymes	Methanocella paludicola SANAE	hypothetical protein
ADP-ribose pyrophosphatase	Methanobacterium sp. SWAN-1	NUDIX hydrolase
OD1-DSC5		
Heme/copper-type cytochrome/quinol oxidases, subunit 2	Ferroglobus placidus DSM 10642	cytochrome C oxidase subunit II
Glycosyltransferase	Methanofollis liminatans	glycosyl transferase group 1
ADP-ribose pyrophosphatase	Methanoculleus bourgensis MS2	Nucleoside triphosphatase
ADP-ribose pyrophosphatase	Methanoculleus bourgensis MS2	Nucleoside triphosphatase
Membrane protein involved in the export of O-antigen and teichoic a	Methanobacterium sp. SWAN-1	polysaccharide biosynthesis protein
OD1-DSC6		
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
Heme/copper-type cytochrome/quinol oxidases, subunit 2	Ferroglobus placidus DSM 10642	cytochrome C oxidase subunit II
Glycosyltransferases involved in cell wall biogenesis	Pyrococcus sp. NA2	dolichol-phosphate mannosyltransferase
Glycosyltransferase	Methanosaeta thermophila PT	glycosyl transferase, group 1
Glycosyltransferase	Methanobacterium formicicum	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanococcus maripaludis C5	group 1 glycosyl transferase
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	Methanocaldococcus sp. FS406-22	phosphoenolpyruvate synthase
Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	Candidatus Nanosalinarum sp. J07AB56	phosphoenolpyruvate synthase/pyruvate phosphate dikinase
Membrane protein involved in the export of O-antigen and teichoic a	Methanobacterium sp. SWAN-1	polysaccharide biosynthesis protein
Membrane protein involved in the export of O-antigen and teichoic a	Methanothermococcus okinawensis IH1	polysaccharide biosynthesis protein
OD1-DSC7		
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
Thiol:disulfide interchange protein	Candidatus Nitrosopumilus salaria	cytochrome C biogenesis protein
Glycosyltransferase	Methanobacterium formicicum	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Predicted glycosyltransferases	Methanolobus psychrophilus R15	glycosyltransferase
Glycosyltransferase	Haloquadratum walsbyi DSM 16790	hexosyltransferase, glycosyltransferase
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
Membrane protein involved in the export of O-antigen and teichoic a	Methanocella paludicola SANAE	putative polysaccharide biosynthesis protein
OD1-DSC8		
ADP-ribose pyrophosphatase	Candidatus Nanosalinarum sp. J07AB56	ADP-ribose pyrophosphatase
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
OD1-DSC9		
Peroxiredoxin	Candidatus Nitrosoarchaeum limnia	alkyl hydroperoxide reductase
Glycosyltransferase	Methanotorris igneus Kol 5	family 2 glycosyl transferase
OD1-DSC10		
Glycosyltransferase	Methanosphaerula palustris E1-9c	family 2 glycosyl transferase
Glycosyltransferase	Haloarcula argentinensis	glycogen synthase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	Archaeoglobus veneficus SNP6	phosphoenolpyruvate synthase
Methylase involved in ubiquinone/menaquinone biosynthesis	Methanosphaerula palustris E1-9c	type 11 methyltransferase
OD1-DSC11		
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
OD1-DSC12		
Cytochrome c biogenesis protein	Candidatus Nitrosoarchaeum limnia	cytochrome C biogenesis protein
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome C biogenesis protein
thymidylate kinase	Candidatus Parvarchaeum acidophilus ARMAN-5	dTMP kinase
UvrC Helix-hairpin-helix N-terminal/GIY-YIG catalytic domain/UvrB/(	Methanosphaera stadtmanae DSM 3091	excinuclease ABC subunit C
Thioredoxin domain	Methanomethylovorans hollandica DSM 15978	glutaredoxin-like protein
Predicted glycosyltransferases	Methanocella arvoryzae MRE50	glycosyl transferase family protein
hypothetical protein	Candidatus Nanosalinarum sp. J07AB56	plastocyanin
OD1-DSC13		
Plastocyanin	Cenarchaeum symbiosum A	copper binding protein, plastocyanin/azurin family
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome c biogenesis protein transmembrane region
Cytochrome c biogenesis protein	Methanomassiliicoccus luminyensis	hypothetical protein
Membrane protein involved in the export of O-antigen and teichoic a	Methanosarcina mazei Tuc01	polysaccharide biosynthesis protein

OD1-DSC product name	Top Hit Species EUK	Top Hit Best Blast
OD1-DSC1		
tRNA-dihydrouridine synthase	Saprolegnia diclina VS20	hypothetical protein SDRG_13668
OD1-DSC3		
tRNA-dihydrouridine synthase	Saprolegnia diclina VS20	hypothetical protein SDRG_13668
OD1-DSC4		
ADP-ribose pyrophosphatase	Xenopus (Silurana) tropicalis	nudix (nucleoside diphosphate linked moiety X)-type motif 1
OD1-DSC10		
ADP-ribose pyrophosphatase	Saprolegnia diclina VS20	hypothetical protein

Table 3.3 Phage-like genes found in OD1-DSC genomes
Genes with best BLAST matches to viral sequences are displayed in this table including the product name annotation by IMG, the top hit species and the top hit BLAST match assessed by the DarkHorse analysis.

OD1-DSC product name	Top Hit Species PHAGE	Top Hit Best Blast
OD1-DSC1		
Zn-dependent alcohol dehydrogenases, class III	Synechococcus phage S-SM2 [Myoviridae]	zinc-containing alcohol dehydrogenase superfamily protein
ADP-heptose synthase, bifunctional sugar kinase/adeny	Synechococcus phage S-SM2 [Myoviridae]	putative carbohydrate kinase
OD1-DSC2		
Helix-destabilising protein	Enterobacteria phage M13 [Inoviridae]	helix destabilising protein
Phage major coat protein, Gp8	Enterobacteria phage M13 [Inoviridae]	structural protein
Type II secretory pathway, component PulD	Enterobacteria phage M13 [Inoviridae]	phage assembly protein
hypothetical protein	Enterobacteria phage M13 [Inoviridae]	small hydrophobic protein
Zonular occludens toxin (Zot)	Enterobacteria phage M13 [Inoviridae]	phage assembly protein
phage/plasmid replication protein, gene II/X family	Enterobacteria phage M13 [Inoviridae]	hypothetical protein
Beta-propeller domains of methanol dehydrogenase typ	Enterobacteria phage M13 [Inoviridae]	gene III
OD1-DSC5		
Type II secretory pathway, component PulD	Enterobacteria phage f1 [Inoviridae]	gene IV, partial
Phage replication protein CRI	Enterobacteria phage M13 [Inoviridae]	replication protein
hypothetical protein	Enterobacteria phage M13 [Inoviridae]	phage assembly protein
OD1-DSC7		
Domain of Unknown Function with PDB structure (DUF3	Clostridium phage phiMMP02 [Myoviridae]	ASCH domain protein
OD1-DSC10		
hypothetical protein	Enterobacteria phage M13 [Inoviridae]	small hydrophobic protein
Phage Coat Protein A	Enterobacteria phage M13 [Inoviridae]	Chain A, Crystal Structure Of The N-Terminal Domains Of Bacteriophag
OD1-DSC12		
Deoxycytidylate deaminase	Bacillus phage 0305phi8-36 [Myoviridae]	deoxycytidylate deaminase
Bacteriophage protein GP30	Caulobacter phage CcrColossus [Siphoviridae]	hypothetical protein CcrColossus_gp169
Type II secretory pathway, component PulD	Enterobacteria phage M13 [Inoviridae]	phage assembly protein
OD1-DSC13		
Aspartyl/asparaginyl-tRNA synthetases	Prochlorococcus phage P-SSM7 [Myoviridae]	tRNA ligase
hypothetical protein	Paenibacillus phage PG1 [Siphoviridae]	hypothetical protein PANG_00064

Supplementary Material

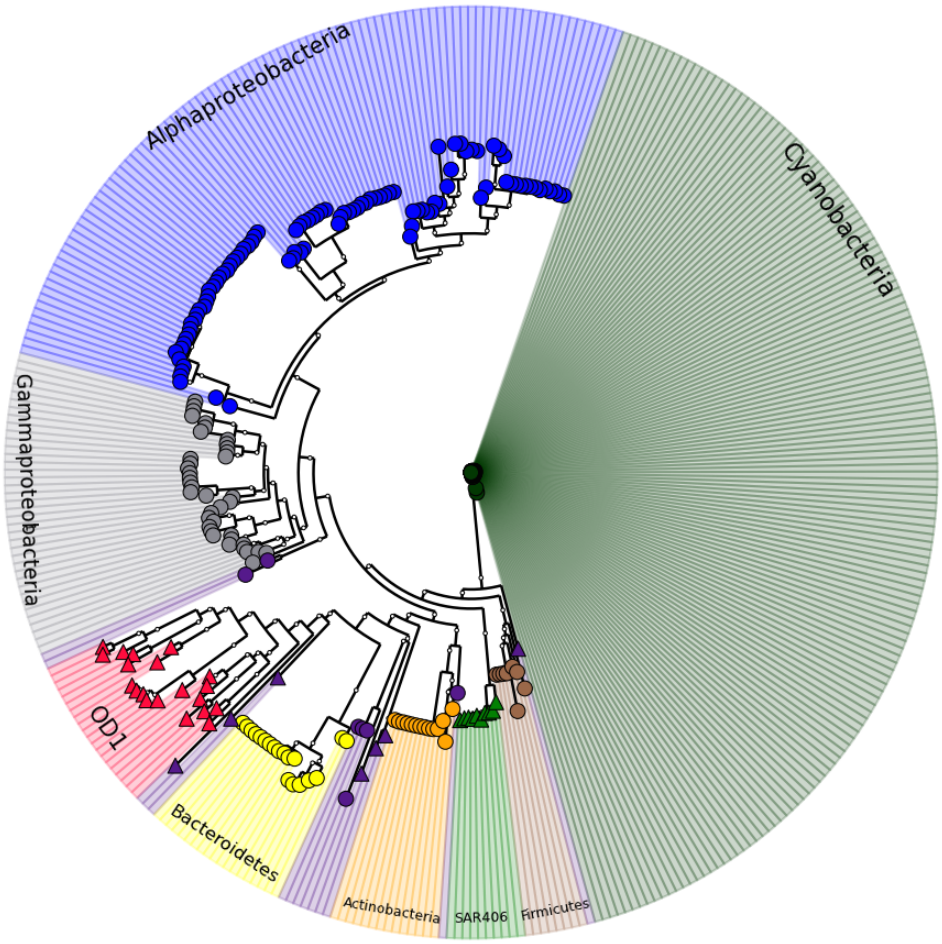

Figure S3.4 - Phylogenetic distribution of Challenger Deep MDAs
Tree shows both the phylogenetic distribution of sorted and successfully amplified samples and their relative abundances. The major players are annotated and colored differently. Of a total of 371, fourteen phyla were represented: *Proteobacteria, Cyanobacteria, Gemmatimonadetes, Firmicutes, Chlamydiae,Actinobacteria, Bacteroidetes*, OP11, JS1, OP3, OD1, BD1-5, TM6, SAR406. The relative abundace distribution is 150 *Cyanobacteria* (40.4%), 97 *Alphaproteobacteria* (26%), 39 *Gammaproteobacteria* (10.5%), 20 OD1 (5.4%), 10 SAR406 (3.7%). All groups less than 2.5% abundance were clustered together and are colored in purple. Circles represent known phyla and triangles represent candidate phyla.

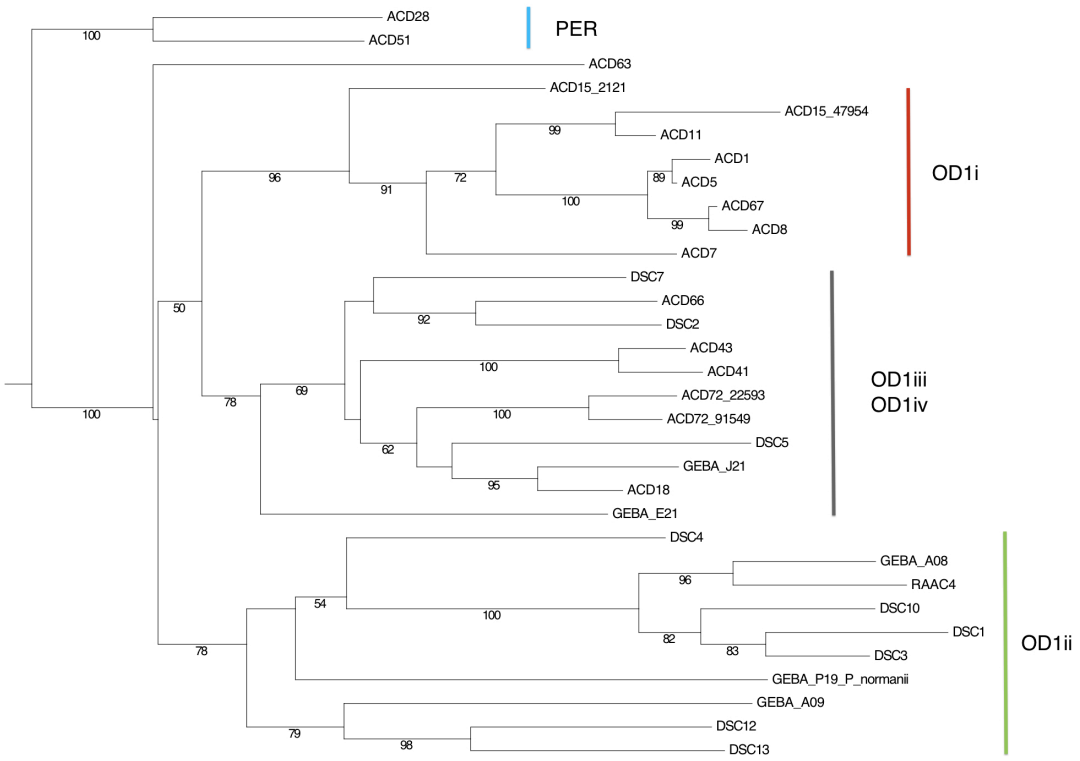

Figure S3.5 Recombinase A phylogenetic distribution of the OD1-DSC genomes
Tree shows the phylogenetic relationships between OD1 genomes of protein Recombinase A encoded by single copy marker gene *recA*. Three distinct clades are distinguishable: OD1i, OD1ii, and a third consisting of OD1iii and OD1iv. Genomes DSC2, 5, 6, and 7 show wandering relationships therefore in these analyses cannot reliably predict the existence of distinct OD1iii and OD1iv clades. ACD genes are from Wrighton *et al*. 2012 and Kantor *et al*. 2013 and were downloaded from http://ggkbase.berkeley.edu/. GEBA genes were dowloaded from https://img.jgi.doe.gov/er/. Whole gene amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are members of the related Peregrines (PER) phylum.

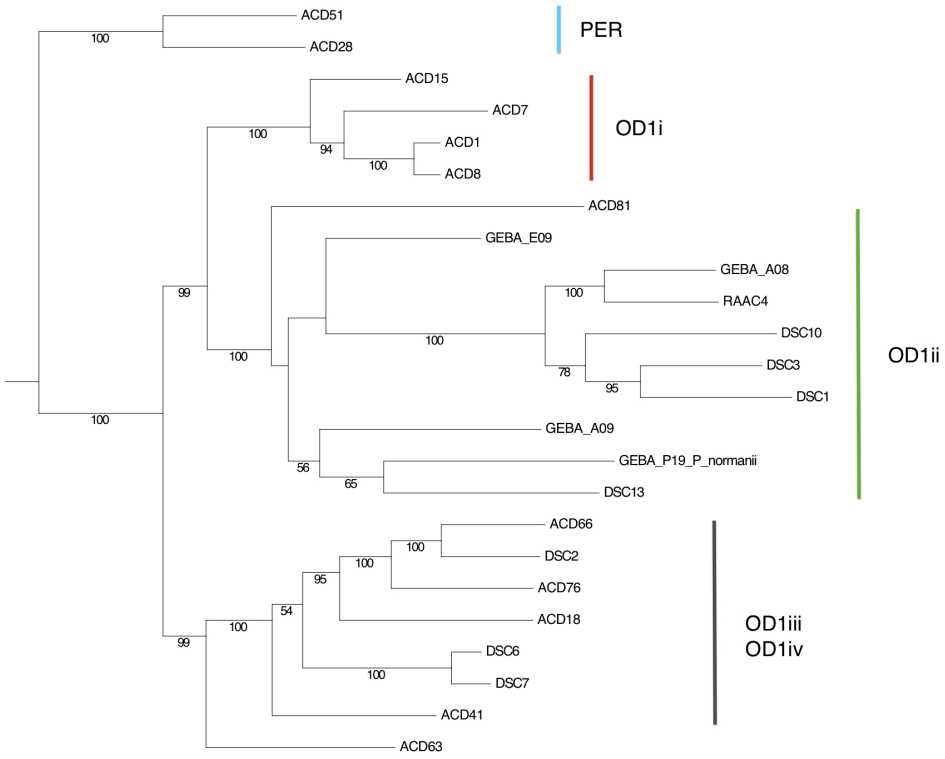

Figure S3.6 RNA Polymerase subunit beta phylogenetic distribution of the OD1-DSC genomes Tree shows the phylogenetic relationships between OD1 genomes of protein RNA Polymerase subunit beta encoded by single copy marker gene *rpoB*. Three distinct clades are distinguishable: OD1i, OD1ii, and a third consisting of OD1iii and OD1iv. Genomes DSC2, 5, 6, and 7 show wandering relationships therefore in these analyses cannot reliably predict the existence of distinct OD1iii and OD1iv clades. ACD genes are from Wrighton *et al*. 2012 and Kantor *et al*. 2013 and were downloaded from http://ggkbase.berkeley.edu/. GEBA genes were dowloaded from https://img.jgi.doe.gov/er/. Whole gene amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are members of the related Peregrines (PER) phylum.

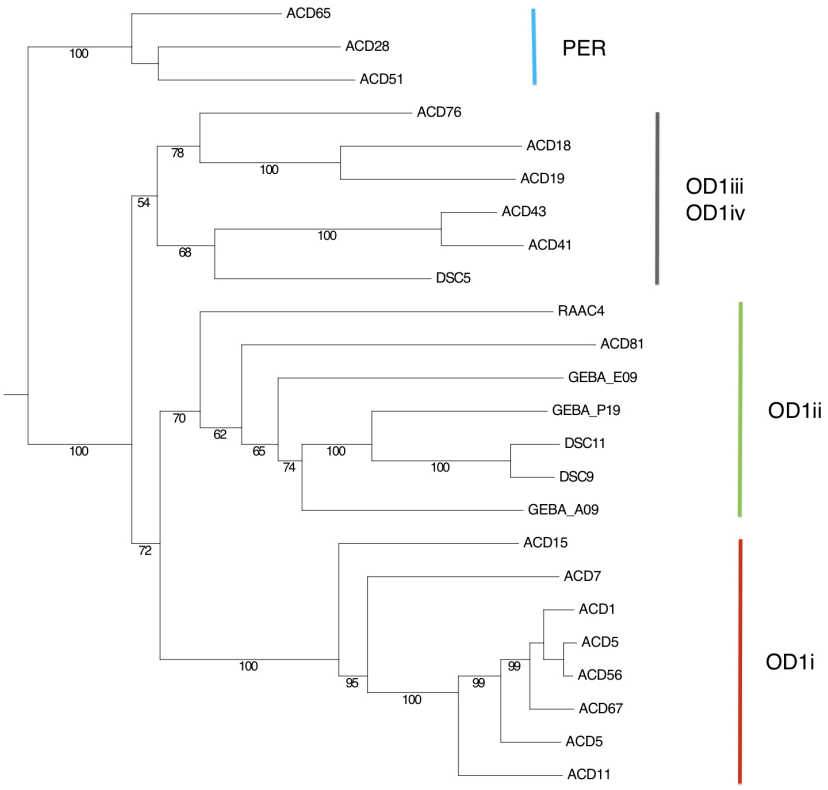

Figure S3.7 DNA Gyrase subunit beta subunit beta phylogenetic distribution of the OD1-DSC genomes

Tree shows the phylogenetic relationships between OD1 genomes of protein DNA Gyrase subunit beta encoded by single copy marker gene. Three distinct clades are distinguishable: OD1i, OD1ii, and a third consisting of OD1iii and OD1iv. Genomes DSC2, 5, 6, and 7 show wandering relationships therefore in these analyses cannot reliably predict the existence of distinct OD1iii and OD1iv clades. ACD genes are from Wrighton *et al*, 2012 and Kantor *et al*, 2013 and were downloaded from http://ggkbase.berkeley.edu/. GEBA genes were dowloaded from https://img.jgi.doe.gov/er/. Whole gene amino acid sequences were aligned with Muscle v.3.8.31, run through ProtTest, and trees created using RAxML v8.0 with the PROTCAT setting and 1000 bootstrapped replicates. Outgroups are members of the related Peregrines (PER) phylum.

Table S3.4 Metabolic potential of OD1-DSC genomes

Metabolic properties highlighted in Figure 3.3 are presented here. Enzyme and protein number and product name can be found as well as the count of how many genomes they are present in and which genomes encode for them. Besides metabolic pathways there are a complete list of transporters present in the genomes.

	Enzymes	EC num	Sum	OD1-DSC1	OD1-DSC 2	OD1-DSC3	OD1-DSC4	OD1-DSC5	OD1-DSC6	OD1-DSC7	OD1-DSC8	OD1-DSC9	OD1-DSC10	OD1-DSC11	OD1-DSC12	OD1-DSC13
	Carbohydrate metabolism															
1 glk	Glucokinase	2.7.1.2	4	1		1				1				1		
2 Pgi/pmi	Bifunctional phosphoglucose/phosphomannose isomera	5.3.1.9; 5.3.1.8	8				1		1	1	1	1		1	1	1
3 pfkA	Phosphofructokinase	2.7.1.11	0													
4 ALDO	Fructose-bisphosphate aldolase class I	4.1.2.13	2					1						1		
4 fbaA	Fructose-bisphosphate aldolase class II	4.1.2.14	3	1											1	1
5 tpiA	Triosephosphate isomerase	5.3.1.1	6	1	1	1			1					1	1	
6 gapA	Glyceraldehyde-3-phosphate dehydrogenase, type I	1.2.1.12	7	1					1	1		1	1	1	1	
7 pgk	Phosphoglycerate kinase	2.7.2.3	7	1		1	1	1				1		1	1	
8 gpmA	2,3-bisphosphoglycerate-dependent phosphoglycerate	5.4.2.1	0													
8 gpmI	2,3-bisphosphoglycerate-independent phosphoglycerat	5.4.2.1	3					1	1							1
8 gpmB	probable phosphoglycerate mutase	5.4.2.2	10	1		1	1	1	1	1		1		1	1	1
9 eno	Enolase (Phosphopyruvate hydratase)	4.2.1.11	3					1	1							1
10 ppdK	Pyruvate phosphate dikinase	2.7.9.1	0													
10 pyk	pyruvate kinase	2.7.1.40	3		1				1						1	
11 pdhA	Pyruvate dehydrogenase E1 component, alpha subunit	1.2.4.1	1	1												
11 pdhB	Pyruvate dehydrogenase E1 component, beta subunit	1.2.4.2	3	1			1								1	
12 pdh	Dihydrolipoamide acetyltransferase E2 component of p	2.3.1.12	0													
13 lpd	Dihydrolipoyl dehydrogenase, E3 component of pyruva	1.8.1.4	1												1	
15 porC	Pyruvate:ferredoxin oxidoreductase, gamma subunit	1.2.7.1	0													
16 porB	Pyruvate:ferredoxin oxidoreductase, beta subunit	1.2.7.1	1					1								
17 porA	Pyruvate:ferredoxin oxidoreductase, alpha subunit	1.2.7.1	2					1								1
18 porD	Pyruvate:ferredoxin oxidoreductase, delta subunit	1.2.7.1	0													
19 fbpase/	Bifunctional Fructose-1-6-bisphosphate aldolase/phosp	4.1.2.13/3.1.3.11	1					1								
20 G6PD, z	glucose-6-phosphate 1-dehydrogenase	1.1.1.49	2				1							1		
21	6-phosphogluconolactonase	3.1.1.31	1											1		
22	6-phosphogluconate dehydrogenase	1.1.1.44	2				1							1		
23 rpiB	Ribose-5-phosphate isomerase B	5.3.1.6	5	1								1	1	1	1	
24 rpe	D-ribulose-5-phosphate 3-epimerase	5.1.3.1	4	1								1		1	1	
25 tktA	Transketolase	2.2.1.1	6	1					1		1	1		1	1	
26 prps	Ribose-phosphate pyrophosphokinase	2.7.6.1	1				1									
27 talA, tal	transaldolase	2.2.1.2	2											1	1	
28 rbks	Ribokinase	2.7.1.15	3									1		1	1	
29 cs	Citrate synthase	2.3.3.1	1						1							
30 acnA	Aconitate hydratase	4.2.1.3	0													
31 idh	Isocitrate dehydrogenase (NADP(+))	1.1.1.42	0													
32 KorA	2-oxoglutarate:ferredoxin oxidoreductase, alpha subun	1.2.7.3	2					1								1
32 korB	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	1.2.7.3	1					1								
32 KorC	2-oxoglutarate:ferredoxin oxidoreductase, gamma sub	1.2.7.4	0													
32 KorD	2-oxoglutarate:ferredoxin oxidoreductase, delta subuni	1.2.7.5	0													
33 sucC	Succinyl-CoA synthetase, ADP-forming , beta subunit	6.2.1.5	0													
33 sucD	Succinyl-CoA synthetase, ADP-forming, alpha subunit	6.2.1.5	2						1	1						
34 sdhA	Succinate dehydrogenase/fumarate reductase, flavopr	1.3.99.1	0													
34 sdhB	Succinate dehydrogenase/fumarate reductase, iron/sul	1.3.99.1	0													
34	Succinate dehydrogenase/fumarate reductase, putative membrane subunit		0													
35 fumB	Fumarate hydratase beta subunit	4.2.1.2	0													
35 fumA	Fumarate hydratase alpha subunit	4.2.1.2	0													
36 mdh	Malate dehydrogenase	1.1.1.37	0													
37 acs	acetyl-coenzyme A synthetase	6.2.1.1	1													1
38	Lactate dehydrogenase and related dehydrogenases	1.1.1.27	5						1	1	1	1		1		
39 adh	alcohol dehydrogenase (NADP+)	1.1.1.2	1									1				
39	Zn-dependent alcohol dehydrogenases, class III	1.1.1.1	1	1												
40	pyruvate decarboxylase	4.1.1.1	0													
41 maeA	Malate dehydrogenase (oxaloacetate- decarboxylating)	EC:1.1.1.38	1	1												

Table S3.4 Metabolic potential of OD1-DSC genomes continued

				OD1-DSC1	OD1-DSC 2	OD1-DSC3	OD1-DSC4	OD1-DSC5	OD1-DSC6	OD1-DSC7	OD1-DSC8	OD1-DSC9	OD1-DSC10	OD1-DSC11	OD1-DSC12	OD1-DSC13
	Respiration															
	Complex I															
	NADH dehydrogenase	1.6.99.3	6	1			1		1					1	1	1
	Complex II															
	Cytochrome bc1 complex, iron-sulfur subunit	1.10.2.2	0													
	Complex III															
	Cytochrome c1		1							1						
	Complex IV															
coxB	Cytochrome c oxidase, subunit II	1.9.3.1	2						1	1						
	Complex V (FoF1-type)															
atpB	F-type H+-transporting ATPase subunit a	3.6.3.14	5	1				1		1				1	1	
atpF	F-type H+-transporting ATPase subunit b	3.6.3.14	4	1				1		1				1		
atpE	F-type H+-transporting ATPase subunit c	3.6.3.14	5	1				1		1				1	1	
atpA	F-type H+-transporting ATPase subunit alpha	3.6.3.14	4	1				1		1				1		
atpD	F-type H+-transporting ATPase subunit beta	3.6.3.14	4	1			1			1				1		
atpC	F-type H+-transporting ATPase subunit epsilon	3.6.3.14	4	1			1	1		1						
atpG	F-type H+-transporting ATPase subunit gamma	3.6.3.14	5	1			1	1		1				1		
ntpA	V-type H+-transporting ATPase subunit A	3.6.3.14	1						1							
ntpB	V-type H+-transporting ATPase subunit B	3.6.3.14	1						1							
ntpC	V-type H+-transporting ATPase subunit C	3.6.3.14	1						1							
ntpD	V-type H+-transporting ATPase subunit D	3.6.3.14	1						1							
ntpF	V-type H+-transporting ATPase subunit F	3.6.3.14	1						1							
ntpI	V-type H+-transporting ATPase subunit I	3.6.3.14	1						1							
ntpK	V-type H+-transporting ATPase subunit K	3.6.3.14	1						1							
narG	nitrate reductase 1, alpha subunit [EC:1.7.99.4]	1.7.99.4	1						1							
narH	nitrate reductase 1, beta subunit [EC:1.7.99.4]	1.7.99.4	1						1							
narJ	nitrate reductase 1, delta subunit	1.7.99.4	1						1							
narI	nitrate reductase 1, gamma subunit [EC:1.7.99.4]	1.7.99.4	1						1							
nirD	nitrite reductase (NAD(P)H) small subunit [EC:1.7.1.4]	1.7.1.4	1		1											
nos	Nitrous oxide reductase	1.7.2.4	1						1							
SQO	sulfide-quinone oxidoreductase	1.8.5.4	1											1		
ppa	inorganic pyrophosphatase [EC:3.6.1.1]	3.6.1.1	6	1		1			1	1		1		1		
ppaX	pyrophosphatase PpaX [EC:3.6.1.1]	3.6.1.1	1											1		
hppA	K(+)-stimulated pyrophosphate-energized sodium pum	3.6.1.1	4					1	1	1						1

Table S3.4 Metabolic potential of OD1-DSC genomes continued

		OD1-DSC1	OD1-DSC 2	OD1-DSC3	OD1-DSC4	OD1-DSC5	OD1-DSC6	OD1-DSC7	OD1-DSC8	OD1-DSC9	OD1-DSC10	OD1-DSC11	OD1-DSC12	OD1-DSC13
Transporters														
Predicted divalent heavy-metal cations transporter	6	1							1		1	1	1	1
Cation transport ATPase	6	1		1			1	1				1		1
Kef-type K+ transport systems, membrane components	3				1						1		1	
Mg2+ and Co2+ transporters	13	1		1	1		1	2	2		1	2	1	1
Biopolymer transport protein	1	1												
Phosphate transport regulator (distant homolog of PhoU)	2				1									1
Cation transport ATPase	4	1			1	1							1	
Nitrate/nitrite transporter	3					1	2							
Na+/H+ antiporter NhaD and related arsenite permeases	1						1							
Fur family transcriptional regulator, ferric uptake regulator	4						1	1	1	1				
ABC transporters														
ABC-type antimicrobial peptide transport system, ATPase component	1													1
ABC-type antimicrobial peptide transport system, permease component	6	1	1	1	1		1					1		
ABC-type dipeptide transport system, periplasmic component	7				1	1	1				1	1	1	1
ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	2												1	1
ABC-type dipeptide/oligopeptide/nickel transport systems, permease compor	2	1	1											
ABC-type dipeptide/oligopeptide/nickel transport systems, permease compor	2						2							
ABC-type metal ion transport system, periplasmic component/surface adhesi	9	1	1	1	1			1	1				2	1
ABC-type Mn/Zn transport systems, ATPase component	2		1				1							
ABC-type Mn2+/Zn2+ transport systems, permease components	2		1				1							
ABC-type multidrug transport system, ATPase and permease components	2		1				1							
ABC-type multidrug transport system, ATPase component	1												1	
ABC-type nitrate/sulfonate/bicarbonate transport systems, periplasmic comp	5	1	1										2	1
ABC-type polysaccharide/polyol phosphate export systems, permease compc	3						1			1		1		
ABC-type polysaccharide/polyol phosphate transport system, ATPase compor	1						1							
ABC-type proline/glycine betaine transport systems, periplasmic components	7				1	1	1				1	1	1	1
ABC-type sugar transport system, periplasmic component	2												1	1
ABC-type transport system involved in Fe-S cluster assembly, permease and	6	1		1	1		1	1					1	
ABC-type transport system involved in Fe-S cluster assembly, permease com	4	1	1				1						1	
ATPase components of ABC transporters with duplicated ATPase domains	1						1							
ATPase components of various ABC-type transport systems, contain duplicat	1			1										
Lipopolysaccharide biosynthesis														
Lipid A core - O-antigen ligase and related enzymes	7	1	1		1	1	1				1			1
3-deoxy-manno-octulosonate cytidylyltransferase (CMF 2.7.7.38	2			1							1			
3-deoxy-D-manno-octulosonate 8-phosphate phosphat 3.1.3.45	1										1			
Glycosyltransferases (pimB) 2.4.1.-	3						1	1			1			
lipid A biosynthesis (KDO)2-(lauroyl)-lipid IVA acyltransferase	2	1								1				

Table S3.4 Metabolic potential of OD1-DSC genomes continued

		OD1-DSC1	OD1-DSC 2	OD1-DSC3	OD1-DSC4	OD1-DSC5	OD1-DSC6	OD1-DSC7	OD1-DSC8	OD1-DSC9	OD1-DSC10	OD1-DSC11	OD1-DSC12	OD1-DSC13
Purine/Pyrimidine biosynthesis														
IMP dehydrogenase [EC:1.1.1.205] (E1.1.1.205, guaB)	5						1	1					2	1
dihydroorotate dehydrogenase (fumarate) [EC:1.3.98.1] (pyrD)	4	1	1			1		1						
thioredoxin reductase (NADPH) [EC:1.8.1.9] (E1.8.1.9, trxB)	2		1									1		
ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1] (E1.17.4.1A	8	1		1	1	1	1				1	1	1	
ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1] (E1.17.4.1B,	2	1					1							
thymidylate synthase [EC:2.1.1.45] (E2.1.1.45, thyA)	8	1		1		1		1		1		1	1	1
phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohyd	1						1							
aspartate carbamoyltransferase catalytic subunit [EC:2.1.3.2] (pyrB, PYR2)	2								1		1			
orotate phosphoribosyltransferase [EC:2.4.2.10] (pyrE)	6			1	1			1		1		1	1	
amidophosphoribosyltransferase [EC:2.4.2.14] (purF, PPAT)	3						1					1	1	
thymidine kinase [EC:2.7.1.21] (E2.7.1.21, tdk)	1		1											
pyruvate kinase [EC:2.7.1.40] (PK, pyk)	3		1				1						1	
uridine kinase [EC:2.7.1.48] (E2.7.1.48, udk)	1		1											
adenylate kinase [EC:2.7.4.3] (E2.7.4.3, adk)	9	1	1		2		1					2	1	1
nucleoside-diphosphate kinase [EC:2.7.4.6] (E2.7.4.6, ndk)	8	1	1	1	1		1					1	1	1
guanylate kinase [EC:2.7.4.8] (E2.7.4.8, gmk)	1						1							
cytidylate kinase [EC:2.7.4.14] (cmk)	2		1			1								
ribose-phosphate pyrophosphokinase [EC:2.7.6.1] (PRPS, prsA)	1				1									
GTP pyrophosphokinase [EC:2.7.6.5] (relA)	1				1									
polyribonucleotide nucleotidyltransferase [EC:2.7.7.8] (pnp, PNPT1)	9	1		1		1	1		1	1	1	1		1
dihydroorotase [EC:3.5.2.3] (URA4, pyrC)	4	1		1			1	1						
guanine deaminase [EC:3.5.4.3] (E3.5.4.3, guaD)	1		1											
cytidine deaminase [EC:3.5.4.5] (E3.5.4.5, cdd)	1							1						
dCMP deaminase [EC:3.5.4.12] (comEB)	4					1		1					1	1
dCTP deaminase [EC:3.5.4.13] (E3.5.4.13, dcd)	5	1								2		2		
inosine triphosphate pyrophosphatase [EC:3.6.1.19] (ITPA)	2				1									1
dUTP pyrophosphatase [EC:3.6.1.23] (E3.6.1.23, dut)	4					1		1					1	1
exopolyphosphatase / guanosine-5'-triphosphate,3'-diphosphate pyrophosph	1						1							
phosphoribosylaminoimidazole carboxylase / phosphoribosylaminoimidazole-	1						1							
orotidine-5'-phosphate decarboxylase [EC:4.1.1.23] (pyrF)	5			1	1		1	1					1	
adenylosuccinate lyase [EC:4.3.2.2] (purB, ADSL)	2						1							1
phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6] (p	1					1								
CTP synthase [EC:6.3.4.2] (E6.3.4.2, pyrG)	7	1		1			1	1	1			1		1
adenylosuccinate synthase [EC:6.3.4.4] (E6.3.4.4, purA)	3							1					1	1
phosphoribosylamine--glycine ligase [EC:6.3.4.13] (purD)	3	1			1							1		
GMP synthase (glutamine-hydrolysing) [EC:6.3.5.2] (E6.3.5.2, guaA)	2						1	1						
carbamoyl-phosphate synthase large subunit [EC:6.3.5.5] (carB, CPA2)	1						1							
DNA polymerase I [EC:2.7.7.7] (DPO1, polA)	5	1		1			1					1		1
DNA polymerase III subunit alpha [EC:2.7.7.7] (DPO3A1, dnaE)	9	1	1		1		1	1			1	1	1	1
DNA polymerase III subunit beta [EC:2.7.7.7] (DPO3B, dnaN)	6	1		1		1	1			1				1
DNA polymerase III subunit delta [EC:2.7.7.7] (DPO3D1, holA)	1									1				
DNA polymerase III subunit delta' [EC:2.7.7.7] (DPO3D2, holB)	6						1	1		1	1		1	1
DNA polymerase III subunit epsilon [EC:2.7.7.7] (DPO3E, dnaQ)	2	1		1										
DNA polymerase III subunit gamma/tau [EC:2.7.7.7] (DPO3G, dnaX)	3		1				1	1						
dITP/XTP pyrophosphatase [EC:3.6.1.19] (rdgB)	2								1			1		
DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6] (rpoA)	9	1		1	1		1	1			1	1	1	1
DNA-directed RNA polymerase subunit beta [EC:2.7.7.6] (rpoB)	7	1	1	1			1	1			1			1
DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6] (rpoC)	8	1	1	1	1		1	1			1			1
5'-nucleotidase [EC:3.1.3.5] (yfbR)	1											1		
uridylate kinase [EC:2.7.4.22] (pyrH)	5	1					1	1	1			1		
phosphoribosylglycinamide formyltransferase 1 [EC:2.1.2.2] (purN)	1				1									
phosphomannomutase / phosphoglucomutase [EC:5.4.2.8 5.4.2.2] (pmm-pg	3				1								1	1

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes - complete

Genes with best BLAST matches to archaea or eukarya are displayed in this table including the product name annotation by IMG, the top hit species and the top hit BLAST match assessed by the DarkHorse

OD1-DSC1		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
ABC-type polysaccharide/polyol phosphate export syste	Haloquadratum sp. J07HQX50	ABC-type polysaccharide/polyol phosphate export syste
Carbamoylphosphate synthase large subunit (split gene	Methanoculleus marisnigri JR1	ATP-grasp enzyme-like protein
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome c biogenesis protein transmembrane regior
Thiol:disulfide interchange protein	Candidatus Nitrososphaera gargensis Ga9.2	disulfide bond oxidoreductase D family protein
hypothetical protein	Methanoplanus limicola	DNA polymerase beta domain protein region
exodeoxyribonuclease III	Methanobacterium sp. SWAN-1	exodeoxyribonuclease III
DNA polymerase III, epsilon subunit and related 3'-5' e	Candidatus Parvarchaeum acidiphilum ARMAN-4	Exonuclease RNase T and DNA polymerase III
Geranylgeranyl pyrophosphate synthase	Methanocella conradii HZ254	geranylgeranyl pyrophosphate synthase
Glucose-6-phosphate isomerase	Thermoplasmatales archaeon SCGC AB-539-C06	glucose-6-phosphate isomerase
Alkyl hydroperoxide reductase, large subunit	Thermococcus sibiricus MM 739	Glutaredoxin/thioredoxin-like protein
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
Glycosyltransferase	Methanothermococcus okinawensis IH1	group 1 glycosyl transferase
hypothetical protein	Halococcus hamelinensis	hypothetical protein
SipW-cognate class signal peptide	Halorubrum sp. T3	hypothetical protein
Uncharacterized membrane protein	Sulfolobales archaeon Acd1	hypothetical protein
hypothetical protein	Archaeoglobus fulgidus DSM 4304	hypothetical protein
parallel beta-helix repeat (two copies)	Methanococcus maripaludis X1	hypothetical protein
Guanosine polyphosphate pyrophosphohydrolases/synt	Halovivax ruber XH-70	hypothetical protein
hypothetical protein	halophilic archaeon J07HB67	hypothetical protein
hypothetical protein	Methanolobus psychrophilus R15	hypothetical protein
uracil-DNA glycosylase, family 4	Pyrococcus sp. ST04	hypothetical protein
Inorganic pyrophosphatase	Methanosphaera stadtmanae DSM 3091	inorganic pyrophosphatase
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
hypothetical protein	Candidatus Nitrosopumilus salaria	membrane protein
hypothetical protein	Halovivax ruber XH-70	nitroreductase family protein
Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclop	Methanosaeta harundinacea 6Ac	peptidyl-prolyl cis-trans isomerase
Plastocyanin	Candidatus Nanosalina sp. J07AB43	plastocyanin
hypothetical protein	Thermoplasmatales archaeon SCGC AB-539-C06	Protein containing DUF1628
Methyltransferase domain	Methanococcus vannielii SB	type 11 methyltransferase
Methyltransferase domain	Methanothermococcus okinawensis IH1	type 12 methyltransferase
hypothetical protein	Candidatus Haloredivivus sp. G17	xenobiotic-transporting ATPase
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
NitCGTAGA464_00618 Thrombospondin type 3 repeat	Anopheles darlingi	hypothetical protein AND_17370
NitCGTAGA464_00921 HhH-GPD superfamily base exci	Candida maltosa Xu316	hypothetical protein G210_3496
NitCGTAGA464_01064 Ribonucleotide reductase, alpha	Pseudogymnoascus destructans 20631-21	hypothetical protein GMDG_08870, partial
NitCGTAGA464_00540 Phosphoribosylaminoimidazoles	Helobdella robusta	hypothetical protein HELRODRAFT_186294
NitCGTAGA464_01028 ABC-type transport system invo	Moniliophthora perniciosa FA553	hypothetical protein MPER_01800
NitCGTAGA464_00597 Uncharacterized conserved prot	Phaseolus vulgaris	hypothetical protein PHAVU_009G221900g
NitCGTAGA464_00936 tRNA-dihydrouridine synthase	Saprolegnia diclina VS20	hypothetical protein SDRG_13668
OD1-DSC2		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Predicted permeases	Methanosarcina barkeri str. Fusaro	conserved hypothetical protein
Geranylgeranyl pyrophosphate synthase	Archaeoglobus profundus DSM 5631	dimethylallyltranstransferase
Glucose-6-phosphate isomerase	Thermoplasmatales archaeon SCGC AB-539-C06	glucose-6-phosphate isomerase
Glutamate dehydrogenase/leucine dehydrogenase	Pyrococcus sp. ST04	glutamate dehydrogenase
hypothetical protein	Thermoplasmatales archaeon SCGC AB-539-C06	hypothetical protein
Cytotoxic translational repressor of toxin-antitoxin stab	Methanosarcina acetivorans C2A	hypothetical protein MA0049
hypothetical protein	Methanococcoides burtonii DSM 6242	hypothetical protein Mbur_0332
hypothetical protein	Methanohalophilus mahii DSM 5219	hypothetical protein Mmah_0573
Uncharacterized conserved protein	Methanococcus maripaludis C7	hypothetical protein MmarC7_0035
ADP-ribose pyrophosphatase	Methanocaldococcus fervens AG86	NUDIX hydrolase
Ion channel	Methanolobus psychrophilus R15	potassium channel protein
Kef-type K+ transport systems, membrane component	Candidatus Haloredivivus sp. G17	sodium/hydrogen exchanger
transporter, CPA2 family (TC 2.A.37)	Methanococcus maripaludis C7	sodium/hydrogen exchanger
Predicted transcriptional regulators	Candidatus Micrarchaeum acidiphilum ARMAN-2	transcriptional regulator, TrmB
uridine kinase (EC 2.7.1.48)	Candidatus Halobonum tyrrellensis	uridine/cytidine kinase
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
Hemolysins and related proteins containing CBS domai	Phaeodactylum tricornutum CCAP 1055/1	predicted protein

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes -complete continued

OD1-DSC3		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
DNA-3-methyladenine glycosylase I (EC 3.2.2.20)	Methanolobus psychrophilus R15	DNA-3-methyladenine glycosylase I
exodeoxyribonuclease III	Methanolobus psychrophilus R15	exodeoxyribonuclease III
DNA polymerase III, epsilon subunit and related 3'-5' e	Candidatus Parvarchaeum acidiphilum ARMAN-4	Exonuclease RNase T and DNA polymerase III
Glucose-6-phosphate isomerase	Thermoplasmatales archaeon SCGC AB-539-C06	glucose-6-phosphate isomerase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanocella conradii HZ254	glycosyltransferase
Electron transfer DM13	Candidatus Nitrosoarchaeum koreensis	hypothetical protein
hypothetical protein	Halogranum salarium	hypothetical protein
hypothetical protein	Thermococcus sp. CL1	hypothetical protein
Lipid A core - O-antigen ligase and related enzymes	Methanocella paludicola SANAE	hypothetical protein
Predicted membrane protein	Methanohalobium evestigatum Z-7303	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
Predicted flavoprotein	Methanosaeta harundinacea 6Ac	hypothetical protein
hypothetical protein	Thermofilum sp. 1910b	hypothetical protein
uracil-DNA glycosylase, family 4	Pyrococcus sp. ST04	hypothetical protein
Predicted membrane protein	Thermococcus litoralis DSM 5473	membrane protein
methionine-S-sulfoxide reductase	Candidatus Nitrosopumilus salaria	methionine sulfoxide reductase A
ADP-ribose pyrophosphatase	Methanobacterium sp. SWAN-1	NUDIX hydrolase
hypothetical protein	Candidatus Nanosalina sp. J07AB43	plastocyanin
Subtilisin-like serine proteases	Methanolobus psychrophilus R15	subtilisin
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
hypothetical protein	Ostreococcus tauri	GTP-binding protein LepA homolog (ISS), partial
hypothetical protein	Nematostella vectensis	hypothetical protein NEMVEDRAFT_v1g224967
tRNA-dihydrouridine synthase	Saprolegnia diclina VS20	hypothetical protein SDRG_13668
ribosomal protein L13, bacterial type	Schizosaccharomyces pombe 972h-	mitochondrial ribosomal protein subunit L13
OD1-DSC4		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
6-phosphogluconate dehydrogenase (decarboxylating)	Candidatus Nanosalinarum sp. J07AB56	6-phosphogluconate dehydrogenase, decarboxylating
hypothetical protein	Methanobacterium formicicum	carbohydrate binding family 6
FOG: PKD repeat	Methanolobus psychrophilus R15	cell surface protein
DnaJ-class molecular chaperone with C-terminal Zn fin	Methanosaeta harundinacea 6Ac	Chaperone protein DnaJ
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
cytidyltransferase-like domain	Methanococcus aeolicus Nankai-3	cytidyltransferase-like protein
Nucleotidyltransferase/DNA polymerase involved in DN	Methanosarcina mazei Tuc01	DNA polymerase IV
hypothetical protein	Fervidicoccus fontis Kam940	DNA-(apurinic or apyrimidinic site) lyase
dTDP-4-dehydrorhamnose 3,5-epimerase and related e	Methanobacterium sp. SWAN-1	dTDP-4-dehydrorhamnose 3,5 epimerase
dTDP-4-dehydrorhamnose 3,5-epimerase	Candidatus Nitrosoarchaeum limnia	dTDP-4-dehydrorhamnose 3,5 epimerase
dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	Methanoculleus marisnigri JR1	dTDP-4-dehydrorhamnose reductase
dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	Thermococcus onnurineus NA1	dTDP-4-dehydrorhamnose reductase
glucose-1-phosphate thymidylylransferase, long form	Methanocella conradii HZ254	glucose-1-phosphate thymidylyltransferase
Predicted glycosyltransferases	Methanocella arvoryzae MRE50	glycosyl transferase family protein
CxxC-x17-CxxC domain	Candidatus Nitrosoarchaeum limnia	hypothetical protein
tRNA threonylcarbamoyl adenosine modification protei	Hyperthermus butylicus DSM 5456	hypothetical protein
hypothetical protein	Methanococcoides burtonii DSM 6242	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
hypothetical protein	Methanohalophilus mahii DSM 5219	hypothetical protein
hypothetical protein	Thermococcus sibiricus MM 739	hypothetical protein
PQQ-like domain	Methanothermobacter thermautotrophicus CaT2	kinase
Major Facilitator Superfamily	Candidatus Nanosalinarum sp. J07AB56	major facilitator superfamily
ABC-type sugar transport system, periplasmic compone	Pyrococcus sp. ST04	MalE-like maltose/maltodextrin ABC transporter
nucleoside diphosphate kinase (EC 2.7.4.6)	Thermoproteus uzoniensis 768-20	nucleoside-diphosphate kinase
PAS domain S-box	Candidatus Nitrosoarchaeum limnia	PAS/PAC sensor signal transduction histidine kinase
Predicted phosphoribosyltransferases	Acidilobus saccharovorans 345-15	Purine phosphoribosyltransferase
Predicted sugar nucleotidyltransferases	uncultured marine crenarchaeote HF4000_ANIW133O4	putative CDP-alcohol phosphatidyltransferase
FOG: WD40-like repeat	Salinarchaeum sp. Harcht-Bsk1	pyrrolo-quinoline quinone
transporter, CPA2 family (TC 2.A.37)	Methanothermococcus okinawensis IH1	sodium/hydrogen exchanger
hypothetical protein	Thermoplasma volcanium GSS1	TVG1559742
Micrococcal nuclease (thermonuclease) homologs	Thermoplasmatales archaeon SCGC AB-540-F20	WD40-like repeat-containing protein
Zn-dependent proteases	Haloquadratum sp. J07HQX50	Zn-dependent protease
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
DnaJ-class molecular chaperone with C-terminal Zn fin	Saprolegnia diclina VS20	DnaJ like subfamily B member 6
Hemolysins and related proteins containing CBS domai	Coccomyxa subellipsoidea C-169	DUF21-domain-containing protein
HrpA-like helicases	Branchiostoma floridae	hypothetical protein
Glycerol uptake facilitator and related permeases (Majo	Capitella teleta	hypothetical protein
ADP-ribose pyrophosphatase	Xenopus (Silurana) tropicalis	nudix (nucleoside diphosphate linked moiety X)-type m
Acyl-CoA dehydrogenases	Latimeria chalumnae	PREDICTED: glutaryl-CoA dehydrogenase, mitochondria
F0F1-type ATP synthase, alpha subunit	Sclerotinia borealis F-4157	putative ATP synthase subunit alpha, mitochondrial

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes -complete continued

OD1-DSC5		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Argininosuccinate synthase	Pyrococcus sp. ST04	argininosuccinate synthase
argininosuccinate synthase (EC 6.3.4.5)	Pyrococcus sp. ST04	argininosuccinate synthase
Aspartyl/asparaginyl-tRNA synthetases	Methanocella conradii HZ254	asparaginyl-tRNA synthetase
cob(I)alamin adenosyltransferase	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
Heme/copper-type cytochrome/quinol oxidases, subuni	Ferroglobus placidus DSM 10642	cytochrome C oxidase subunit II
Glycosyltransferase	Methanofollis liminatans	glycosyl transferase group 1
Predicted acetyltransferase	Candidatus Nitrosopumilus salaria	histone acetyltransferase
Histone acetyltransferase	Methanothermobacter marburgensis str. Marburg	histone acetyltransferase
AAA domain	Thermofilum sp. 1910b	hypothetical protein
hypothetical protein	Haloquadratum walsbyi	hypothetical protein
hypothetical protein	Thermoplasmatales archaeon SCGC AB-539-C06	hypothetical protein
hypothetical protein	Methanohalobium evestigatum Z-7303	hypothetical protein
hypothetical protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
ADP-ribose pyrophosphatase	Methanoculleus bourgensis MS2	Nucleoside triphosphatase
ADP-ribose pyrophosphatase	Methanoculleus bourgensis MS2	Nucleoside triphosphatase
OD1CGTACT464_00430 PD-(D/E)XK nuclease superfan	Thermoplasmatales archaeon SCGC AB-540-F20	PD-(D/E)XK nuclease superfamily
Membrane protein involved in the export of O-antigen a	Methanobacterium sp. SWAN-1	polysaccharide biosynthesis protein
ribonucleoside-diphosphate reductase class II (EC 1.17	Methanohalobium evestigatum Z-7303	ribonucleoside-diphosphate reductase
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
Metal-dependent hydrolase	Pneumocystis murina B123	hypothetical protein
UvrC Helix-hairpin-helix N-terminal	Rhodiola fastigiata	excinuclease ABC subunit C
Calcineurin-like phosphoesterase	Batrachochytrium dendrobatidis JAM81	hypothetical protein
hypothetical protein	Coccomyxa subellipsoidea C-169	hypothetical protein
ATPase components of ABC transporters with duplicate	Lolium perenne	putative iron inhibited ABC transporter 2, partial
OD1-DSC6		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
ATP:cob(I)alamin adenosyltransferase	halophilic archaeon J07HX64	ATP:cob(I)alamin adenosyltransferase
Uncharacterized protein involved in tolerance to divaler	Archaeoglobus profundus DSM 5631	CutA1 divalent ion tolerance protein
Cytidylyltransferase	Thermoplasmatales archaeon SCGC AB-539-N05	cytidyltransferase-related enzyme
Heme/copper-type cytochrome/quinol oxidases, subuni	Ferroglobus placidus DSM 10642	cytochrome C oxidase subunit II
Glycosyltransferases involved in cell wall biogenesis	Pyrococcus sp. NA2	dolichol-phosphate mannosyltransferase
RmlD substrate binding domain	Methanosarcina acetivorans C2A	dTDP-4-dehydrorhamnose reductase
Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.2	Methanosarcina mazei Tuc01	Glucose-1-phosphate thymidylyltransferase
Glycosyltransferase	Methanosaeta thermophila PT	glycosyl transferase, group 1
Glycosyltransferase	Methanobacterium formicicum	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanococcus maripaludis C5	group 1 glycosyl transferase
Kunitz/Bovine pancreatic trypsin inhibitor domain	Candidatus Nitrosoarchaeum koreensis	hypothetical protein
hypothetical protein	Halococcus morrhuae	hypothetical protein
hypothetical protein	Halogranum salarium	hypothetical protein
hypothetical protein	Aciduliprofundum sp. MAR08-339	hypothetical protein
Prenyltransferase, beta subunit	Archaeoglobus sulfaticallidus PM70-1	hypothetical protein
hypothetical protein	Cenarchaeum symbiosum A	hypothetical protein
hypothetical protein	Methanoculleus marisnigri JR1	hypothetical protein
Uncharacterized conserved protein	Methanohalobium evestigatum Z-7303	hypothetical protein
Uncharacterized conserved protein	Methanothermococcus okinawensis IH1	hypothetical protein
hypothetical protein	Methanohalophilus mahii DSM 5219	hypothetical protein
Glycosyltransferase	Pyrococcus horikoshii OT3	hypothetical protein
Inorganic pyrophosphatase	Sulfolobales archaeon Acd1	inorganic pyrophosphatase
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
Nucleoside-diphosphate-sugar pyrophosphorylase invol	Pyrococcus sp. ST04	mannose-1-phosphate guanyltransferase
nucleoside diphosphate kinase (EC 2.7.4.6)	Thermoproteus uzoniensis 768-20	nucleoside-diphosphate kinase
Phosphoenolpyruvate synthase/pyruvate phosphate dik	Methanocaldococcus sp. FS406-22	phosphoenolpyruvate synthase
Phosphoenolpyruvate synthase/pyruvate phosphate dik	Candidatus Nanosalinarum sp. J07AB56	phosphoenolpyruvate synthase/pyruvate phosphate dik
Phosphohistidine swiveling domain	Candidatus Nanosalinarum sp. J07AB56	phosphoenolpyruvate synthase/pyruvate phosphate dik
ABC-type polysaccharide/polyol phosphate export syste	Methanosaeta harundinacea 6Ac	Polysaccharide ABC transporter, permease protein
Membrane protein involved in the export of O-antigen a	Methanobacterium sp. SWAN-1	polysaccharide biosynthesis protein
Membrane protein involved in the export of O-antigen a	Methanothermococcus okinawensis IH1	polysaccharide biosynthesis protein
Predicted hydrolases or acyltransferases (alpha/beta hy	Candidatus Micrarchaeum acidiphilum ARMAN-2	Protein of unknown function DUF1749
pseudaminic acid synthase	Methanobrevibacter smithii CAG:186	pseudaminic acid synthase
Predicted deacylase	Thermoplasmatales archaeon SCGC AB-539-N05	putative deacylase
Predicted glycosyltransferases	Methanocella paludicola SANAE	putative glycosyltransferase
HD superfamily phosphohydrolases	Candidatus Haloredivivus sp. G17	putative metal-dependent phosphohydrolase
Predicted membrane protein	Methanolobus psychrophilus R15	putative small multi-drug export protein
Ribonuclease HI	Candidatus Caldiarchaeum subterraneum	ribonuclease HI
Methylase involved in ubiquinone/menaquinone biosynt	Methanothermococcus okinawensis IH1	type 12 methyltransferase
Excinuclease ABC subunit C	Methanoculleus bourgensis MS2	UvrABC system protein C
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
Queuine/archaeosine tRNA-ribosyltransferase	Pavlova lutheri	chloroplast queuine tRNA ribosyltransferase
Phosphoribosylaminoimidazolesuccinocarboxamide (SA	Trichoplax adhaerens	hypothetical protein TRIADDRAFT_56870
Predicted hydrolases of HD superfamily	Albugo laibachii Nc14	PREDICTED: hypothetical protein isoform 1
NADH:ubiquinone oxidoreductase 49 kD subunit 7	Hydra vulgaris	PREDICTED: NADH dehydrogenase
Orotate phosphoribosyltransferase	Trypanosoma congolense IL3000	putative orotidine-5-phosphate decarboxylase/orotate p

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes -complete continued

OD1-DSC7		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
hypothetical protein	Candidatus Nanosalina sp. J07AB43	AhpC/TSA family protein
AhpC/TSA family	Halonotius sp. J07HN6	AhpC/TSA family protein
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
cytidyltransferase-like domain	Thermoplasmatales archaeon SCGC AB-539-N05	cytidyltransferase-related enzyme
Thiol:disulfide interchange protein	Candidatus Nitrosopumilus salaria	cytochrome C biogenesis protein
NTP pyrophosphohydrolases including oxidative damag	Candidatus Nitrosopumilus salaria	diadenosine 55-P1,P4-tetraphosphate pyrophosphohyd
Uncharacterized protein involved in tolerance to divaler	Pyrococcus sp. ST04	divalent cation tolerance protein
dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	Methanocella arvoryzae MRE50	dTDP-4-dehydrorhamnose reductase
OD1TAATAT464_00362 Glucose-1-phosphate thymidyly	Methanosarcina mazei Tuc01	Glucose-1-phosphate thymidylyltransferase
Glycosyltransferase	Methanobacterium formicicum	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Predicted glycosyltransferases	Methanolobus psychrophilus R15	glycosyltransferase
Glycosyltransferase	Haloquadratum walsbyi DSM 16790	hexosyltransferase, glycosyltransferase
OD1TAATAT464_00526 Capsule polysaccharide export	Candidatus Nitrosoarchaeum limnia	hypothetical protein
hypothetical protein	Candidatus Nitrosoarchaeum limnia	hypothetical protein
Highly conserved protein containing a thioredoxin dom	Haladaptatus paucihalophilus	hypothetical protein
ATP-dependent exoDNAse (exonuclease V) beta subuni	Haloquadratum walsbyi	hypothetical protein
hypothetical protein	Methanoculleus bourgensis MS2	hypothetical protein
hypothetical protein	Candidatus Nanosalinarum sp. J07AB56	hypothetical protein
2'-5' RNA ligase superfamily	Methanoculleus marisnigri JR1	hypothetical protein
Uncharacterized conserved protein	Methanobacterium sp. SWAN-1	hypothetical protein
hypothetical protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
Dephospho-CoA kinase	Pyrococcus sp. ST04	hypothetical protein
Aspartate/tyrosine/aromatic aminotransferase	Methanothermococcus okinawensis IH1	LL-diaminopimelate aminotransferase
Glycosyltransferase	Haloarcula argentinensis	LPS biosynthesis RfbU related protein
Uncharacterized membrane protein	Nitrosopumilus sp. AR	magnesium transporter MgtC
Nucleoside-diphosphate-sugar pyrophosphorylase invol	Pyrococcus sp. ST04	mannose-1-phosphate guanyltransferase
N-acetylneuraminate synthase (EC 2.5.1.56)	Methanobrevibacter smithii CAG:186	pseudaminic acid synthase
Endonuclease IV (EC 3.1.21.-)	Methanosaeta harundinacea 6Ac	putative endonuclease 4
Membrane protein involved in the export of O-antigen a	Methanocella paludicola SANAE	putative polysaccharide biosynthesis protein
hypothetical protein	Thermoproteus uzoniensis 768-20	putative RNA methylase
Predicted DNA modification methylase	Pyrococcus yayanosii CH1	RNA methylase
tRNA threonylcarbamoyl adenosine modification protein	Candidatus Nanosalina sp. J07AB43	sua5/YciO/YrdC/YwlC family protein
Succinyl-CoA synthetase, beta subunit	Methanocaldococcus fervens AG86	succinyl-CoA synthetase, beta subunit
Excinuclease ABC subunit C	Methanoculleus bourgensis MS2	UvrABC system protein C [Methanoculleus bourgensis M
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
LSU ribosomal protein L17P	Saprolegnia diclina VS20	50S ribosomal protein L17
hypothetical protein	Neospora caninum Liverpool	hypothetical protein
ATPase family associated with various cellular activities	Sordaria macrospora k-hell	hypothetical protein
OD1TAATAT464_00245 four helix bundle protein	Volvox carteri f. nagariensis	hypothetical protein
hypothetical protein	Micromonas sp. RCC299	predicted protein
Thymidylate kinase	Trichomonas vaginalis G3	thymidylate kinase family protein
Protein of unknown function (DUF933)	Lotus japonicus	unknown
OD1-DSC8		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Predicted 6-phosphogluconate dehydrogenase	Candidatus Nanosalina sp. J07AB43	6-phosphogluconate dehydrogenase, decarboxylating
ADP-ribose pyrophosphatase	Candidatus Nanosalinarum sp. J07AB56	ADP-ribose pyrophosphatase
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
Cysteine sulfinate desulfinase/cysteine desulfurase and	Methanolobus psychrophilus R15	aminotransferase
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
cytidyltransferase-like domain	Candidatus Nitrosoarchaeum koreensis	cytidyltransferase
Thiol:disulfide interchange protein	Candidatus Nitrososphaera gargensis Ga9.2	disulfide bond oxidoreductase D family protein
Nucleotidyltransferase/DNA polymerase involved in DN	Methanoculleus bourgensis MS2	DNA polymerase IV (archaeal DinB-like DNA polymeras
serine hydroxymethyltransferase (EC 2.1.2.1)	Candidatus Micrarchaeum acidiphilum ARMAN-2	Glycine hydroxymethyltransferase
Lactate dehydrogenase and related dehydrogenases	Thermofilum sp. 1910b	glyoxylate reductase
hypothetical protein	Halococcus salifodinae	hypothetical protein
Alpha/beta hydrolase family	Thermofilum sp. 1910b	hypothetical protein
Major Facilitator Superfamily	Candidatus Nanosalinarum sp. J07AB56	major facilitator superfamily
Phosphoenolpyruvate synthase/pyruvate phosphate dik	Candidatus Nanosalinarum sp. J07AB56	phosphoenolpyruvate synthase/pyruvate phosphate dik
Glutaredoxin and related proteins	Methanocaldococcus sp. FS406-22	redox-active disulfide protein 1
uracil-DNA glycosylase, family 4	Thermococcus onnurineus NA1	uracil-DNA glycosylase
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
hypothetical protein	Pan paniscus	PREDICTED: zinc finger protein 714-like
aspartate carbamoyltransferase (EC 2.1.3.2)	Cicer arietinum	PREDICTED: aspartate carbamoyltransferase 3, chlorop

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes -complete continued

OD1-DSC9		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Peroxiredoxin	Candidatus Nitrosoarchaeum limnia	alkyl hydroperoxide reductase
NeoTCCAGA494_00359 Predicted phosphatase/phosph	Sulfolobus acidocaldarius N8	beta-phosphoglucomutase
Glycosyltransferase	Methanotorris igneus Kol 5	family 2 glycosyl transferase
serine hydroxymethyltransferase (EC 2.1.2.1)	Candidatus Micrarchaeum acidiphilum ARMAN-2	Glycine hydroxymethyltransferase
A/G-specific DNA-adenine glycosylase (EC 3.2.2.-)	Methanoculleus marisnigri JR1	HhH-GPD family protein
haloacid dehalogenase superfamily, subfamily IA, varia	Halopiger xanaduensis SH-6	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
Major Facilitator Superfamily	Candidatus Nanosalinarum sp. J07AB56	major facilitator superfamily
Methylated DNA-protein cysteine methyltransferase	Pyrolobus fumarii 1A	methylated-DNA/protein-cysteinemethyltransferase
Subtilisin-like serine proteases	Methanolobus psychrophilus R15	subtilisin
tryptophan synthase, alpha chain (EC 4.2.1.20)	Methanobacterium sp. AL-21	Tryptophan synthase subunit alpha [Methanobacterium
OD1-DSC product name	**Top Hit Species EUK**	
FOG: Ankyrin repeat	Trichomonas vaginalis G3	ankyrin repeat protein
FOG: Ankyrin repeat	Trichomonas vaginalis G3	ankyrin repeat protein
hypothetical protein	Nematostella vectensis	predicted protein
Aldo/keto reductases, related to diketogulonate reducta	Ceratotherium simum simum	PREDICTED: alcohol dehydrogenase [NADP(+)] isoform
FOG: Ankyrin repeat	Amphimedon queenslandica	PREDICTED: ankyrin repeat domain-containing protein
Putative binding domain	Pan troglodytes	PREDICTED: histone demethylase UTY-like
OD1-DSC10		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Predicted carbamoyl transferase, NodU family	Nitrosopumilus maritimus SCM1	carbamoyltransferase
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
hypothetical protein	Haloquadratum walsbyi C23	conserved hypothetical protein
Glycosyltransferase	Methanosphaerula palustris E1-9c	family 2 glycosyl transferase
hypothetical protein	Candidatus Parvarchaeum acidophilus ARMAN-5	Glycogen debranching protein-like protein [Candidatus
Glycosyltransferase	Haloarcula argentinensis	glycogen synthase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
Glycosyltransferase	Haloarcula argentinensis	group 1 glycosyl transferase
hypothetical protein	Candidatus Nitrosopumilus koreensis AR1	HNH endonuclease [Candidatus Nitrosopumilus koreens
Methyltransferase small domain	Methanomassiliicoccus luminyensis	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
Predicted flavoprotein	Methanosaeta harundinacea 6Ac	hypothetical protein
hypothetical protein	Nitrosopumilus maritimus SCM1	hypothetical protein
Nitroreductase	Archaeoglobus veneficus SNP6	nitroreductase
Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclop	Methanosaeta harundinacea 6Ac	peptidyl-prolyl cis-trans isomerase
Phosphoenolpyruvate synthase/pyruvate phosphate dik	Archaeoglobus veneficus SNP6	phosphoenolpyruvate synthase
Phosphohistidine swiveling domain	halophilic archaeon J07HX5	phosphoenolpyruvate synthase/pyruvate phosphate dik
hypothetical protein	Candidatus Nanosalina sp. J07AB43	plastocyanin
hypothetical protein	Candidatus Nitrosopumilus salaria	protein-disulfide isomerase
Predicted esterase of the alpha/beta hydrolase fold	Candidatus Nanosalinarum sp. J07AB56	putative esterase
Glucose/sorbosone dehydrogenases	Methanobacterium sp. SWAN-1	quinoprotein glucose dehydrogenase
Kef-type K+ transport systems, membrane component	Methanocaldococcus vulcanius M7	sodium/hydrogen exchanger
Methylase involved in ubiquinone/menaquinone biosynt	Methanosphaerula palustris E1-9c	type 11 methyltransferase
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
ribosomal protein L13, bacterial type	Albugo laibachii Nc14	50S ribosomal protein L13 putative
aspartate carbamoyltransferase (EC 2.1.3.2) [494_227	Zea mays	aspartate carbamoyltransferase 1
HrpA-like helicases	Nannochloropsis gaditana CCMP526	deah (asp-glu-ala-his) box polypeptide 16, partial
CTP synthase (UTP-ammonia lyase)	Citrus clementina	hypothetical protein
ADP-ribose pyrophosphatase	Saprolegnia diclina VS20	hypothetical protein
hypothetical protein	Paralichthys olivaceus	interferon
HrpA-like helicases	Fomitiporia mediterranea MF3/22	P-loop containing nucleoside triphosphate hydrolase pr

Table S3.5 Horizontally transferred genes from archaeal and eukaryotes best matches for OD1-DSC genomes -complete continued

OD1-DSC11		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Peroxiredoxin	Candidatus Nitrosoarchaeum koreensis	alkyl hydroperoxide reductase
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
hypothetical protein	Natrinema pellirubrum DSM 15624	dipeptidyl aminopeptidase/acylaminoacyl peptidase
Nucleotidyltransferase/DNA polymerase involved in DN	Methanoculleus bourgensis MS2	DNA polymerase IV (archaeal DinB-like DNA polymeras
DNA-3-methyladenine glycosylase I (EC 3.2.2.20)	Methanobacterium sp. SWAN-1	DNA-3-methyladenine glycosylase I
serine hydroxymethyltransferase (EC 2.1.2.1)	Candidatus Micrarchaeum acidiphilum ARMAN-2	Glycine hydroxymethyltransferase
Glucoamylase and related glycosyl hydrolases	Natrinema pellirubrum DSM 15624	glycosyl hydrolase, glucoamylase
hypothetical protein	Candidatus Nitrosopumilus salaria	hypothetical protein
hypothetical protein	Halogranum salarium	hypothetical protein
Uncharacterized membrane protein	Haloferax mediterranei ATCC 33500	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
hypothetical protein	Methanohalophilus mahii DSM 5219	hypothetical protein
hypothetical protein	Vulcanisaeta moutnovskia 768-28	hypothetical protein
Major Facilitator Superfamily	Candidatus Nanosalinarum sp. J07AB56	major facilitator superfamily
O-6-methylguanine DNA methyltransferase	Pyrolobus fumarii 1A	methylated-DNA/protein-cysteinemethyltransferase
orotate phosphoribosyltransferase (EC 2.4.2.10)	Methanosaeta concilii GP6	orotate phosphoribosyltransferase
Predicted membrane protein	Methanolobus psychrophilus R15	putative small multi-drug export protein
Subtilisin-like serine proteases	Methanolobus psychrophilus R15	subtilisin
Subtilisin-like serine proteases	Methanolobus psychrophilus R15	subtilisin
thioredoxin reductase (NADPH) (EC 1.8.1.9)	Sulfolobus islandicus LAL14/1	thioredoxin reductase
EMAP domain	Candidatus Korarchaeum cryptofilum OPF8	tRNA-binding domain-containing protein
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
Predicted DNA-binding protein with PD1-like DNA-bindi	Micromonas pusilla CCMP1545	predicted protein

OD1-DSC12		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
4-aminobutyrate aminotransferase and related aminotr	Thermococcus sp. 4557	4-aminobutyrate aminotransferase
A/G-specific DNA-adenine glycosylase (EC 3.2.2.-)	Methanoregula formicica SMSP	A/G-specific DNA glycosylase
AhpC/TSA family	Candidatus Nanosalina sp. J07AB43	AhpC/TSA family protein
Uncharacterized conserved protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
cystathionine gamma-lyase (EC 4.4.1.1)	Pyrococcus yayanosii CH1	cystathionine gamma-lyase
Cytochrome c biogenesis protein	Candidatus Nitrosoarchaeum limnia	cytochrome C biogenesis protein
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome C biogenesis protein
thymidylate kinase	Candidatus Parvarchaeum acidophilus ARMAN-5	dTMP kinase
UvrC Helix-hairpin-helix N-terminal/GIY-YIG catalytic d	Methanosphaera stadtmanae DSM 3091	excinuclease ABC subunit C
Thioredoxin domain	Methanomethylovorans hollandica DSM 15978	glutaredoxin-like protein
Predicted glycosyltransferases	Methanocella arvoryzae MRE50	glycosyl transferase family protein
Iron-sulfur cluster assembly accessory protein	Nitrosopumilus sp. SJ	heme biosynthesis protein HemY
hypothetical protein	Halogranum salarium	hypothetical protein
hypothetical protein	Halogranum salarium	hypothetical protein
Predicted transcriptional regulators	Thermococcus sp. CL1	hypothetical protein
hypothetical protein	halophilic archaeon J07HB67	hypothetical protein
Lamin Tail Domain	Methanosaeta concilii GP6	hypothetical protein
Predicted flavoprotein	Methanosaeta harundinacea 6Ac	hypothetical protein
hypothetical protein	Methanosaeta thermophila PT	hypothetical protein
SPFH domain, Band 7 family protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
hypothetical protein	Candidatus Micrarchaeum acidiphilum ARMAN-2	hypothetical protein
Sugar phosphate isomerases/epimerases	Haloferax volcanii DS2	isomerase
nucleoside diphosphate kinase (EC 2.7.4.6)	Pyrobaculum sp. 1860	nucleoside diphosphate kinase
hypothetical protein	Candidatus Nanosalinarum sp. J07AB56	plastocyanin
Glucose/sorbosone dehydrogenases	Methanolobus psychrophilus R15	quinoprotein glucose dehydrogenase
transporter, CPA2 family (TC 2.A.37)	Methanothermococcus okinawensis IH1	sodium/hydrogen exchanger
Thioredoxin reductase	uncultured Acidilobus sp. MG	thioredoxin-disulfide reductase
hypothetical protein	Candidatus Nitrosopumilus salaria	thrombospondin
Zn-dependent proteases	Halogeometricum borinquense DSM 11551	zn-dependent protease
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
LSU ribosomal protein L13P	Rhodosporidium toruloides NP11	50S ribosomal protein l13
hypothetical protein	Moniliophthora perniciosa FA553	hypothetical protein
hypothetical protein	Vicugna pacos	PREDICTED: LOW QUALITY PROTEIN: zinc finger protei
Hemolysins and related proteins containing CBS domai	Odobenus rosmarus divergens	PREDICTED: metal transporter CNNM4

OD1-DSC13		
OD1-DSC product name	**Top Hit Species ARCH**	**Top Hit Best Blast**
Uncharacterized bacitracin resistance protein	Methanococcoides burtonii DSM 6242	bacitracin resistance protein BacA
Na+-driven multidrug efflux pump	Thermococcus barophilus MP	capsular polysaccharide biosynthesis protein
hypothetical protein	Candidatus Micrarchaeum acidiphilum ARMAN-2	CMP/dCMP deaminase zinc-binding
Plastocyanin	Cenarchaeum symbiosum A	copper binding protein, plastocyanin/azurin family
Cytochrome c biogenesis protein	Nitrosopumilus maritimus SCM1	cytochrome c biogenesis protein transmembrane regior
diaminopimelate decarboxylase	Fervidicoccus fontis Kam940	diaminopimelate decarboxylase
DNA-3-methyladenine glycosylase I (EC 3.2.2.20)	Methanobacterium sp. AL-21	DNA-3-methyladenine glycosylase I
dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)	Methanoculleus marisnigri JR1	dTDP-4-dehydrorhamnose reductase
thymidylate kinase	Candidatus Parvarchaeum acidophilus ARMAN-5	dTMP kinase
UvrC Helix-hairpin-helix N-terminal	Nitrosopumilus maritimus SCM1	excinuclease ABC subunit C
Carbohydrate binding module (family 6)	Salinarchaeum sp. Harcht-Bsk1	glucan endo-1,3-beta-D-glucosidase
Glutamate dehydrogenase/leucine dehydrogenase	Candidatus Caldiarchaeum subterraneum	glutamate dehydrogenase (NAD(P)+)
Uncharacterized conserved protein	Candidatus Nitrosoarchaeum limnia	hypothetical protein
Cytochrome c biogenesis protein	Methanomassiliicoccus luminyensis	hypothetical protein
Predicted sugar kinase	Natronorubrum tibetense	hypothetical protein
hypothetical protein	Halorhabdus utahensis DSM 12940	hypothetical protein
Protein of unknown function (DUF3179)	Methanohalobium evestigatum Z-7303	hypothetical protein
Uncharacterized conserved protein	Methanococcus maripaludis C7	hypothetical protein
MoxR-like ATPases	Pyrococcus sp. ST04	methanol dehydrogenase regulatory protein
nucleoside diphosphate kinase (EC 2.7.4.6)	Pyrobaculum sp. 1860	nucleoside diphosphate kinase
oligopeptide/dipeptide ABC transporter, ATP-binding pr	Thermofilum sp. 1910b	peptide ABC transporter ATPase
Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclop	Thermoplasmatales archaeon BRNA1	Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclopl
HD superfamily phosphohydrolases	Methanolinea tarda	phosphohydrolase
DNA polymerase III, alpha subunit	Fervidicoccus fontis Kam940	PHP C-terminal domain-containing protein
Membrane protein involved in the export of O-antigen a	Methanosarcina mazei Tuc01	polysaccharide biosynthesis protein
Predicted esterase of the alpha/beta hydrolase fold	Candidatus Micrarchaeum acidiphilum ARMAN-2	protein of unknown function DUF1234
Predicted membrane protein	Halorubrum sp. J07HR59	putative membrane protein
Predicted glycosyltransferases	Pyrococcus sp. ST04	rhamnosyl transferase-like protein
Kef-type K+ transport systems, membrane component	Methanothermococcus okinawensis IH1	sodium/hydrogen exchanger
OD1-DSC product name	**Top Hit Species EUK**	**Top Hit Best Blast**
Spermidine synthase	Candida maltosa Xu316	Spermine synthase
Periplasmic protease	Lolium perenne	carboxyl-terminal-processing protease precursor, partia
Hemolysins and related proteins containing CBS domai	Coccomyxa subellipsoidea C-169	DUF21-domain-containing protein
Uncharacterized conserved protein (DUF2181)	Xenopus (Silurana) tropicalis	PREDICTED: protein FAM151B isoform X1
Aminopeptidase N	Bombus impatiens	PREDICTED: puromycin-sensitive aminopeptidase-like

REFERENCES

Alba BM, Gross CA (2004). Regulation of the Escherichia coliσE-dependent envelope stress response. *Molecular Microbiology* **52:** 613-619.

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* **31:** 533-538.

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

Ayers M, Howell PL, Burrows LL (2010). Architecture of the type II secretion and type IV pilus machineries. *Future Microbiology* **5:** 1203-1218.

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A Lesin, VM, Nikolenko, SI, Pham, S, Prjibelski, AD, Pyshkin, AV, Sirotkin, AV, Vyahhi, N, Tesler, G, Alekseyev, MA, Pevzner, PA (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19:** 455-477.

Bessman MJ, Frick DN, O'Handley SF (1996). The MutT Proteins or "Nudix" Hydrolases, a Family of Versatile, Widely Distributed, "Housecleaning" Enzymes. *Journal of Biological Chemistry* **271:** 25059-25062.

Browning DF, Busby SJW (2004). The regulation of bacterial transcription initiation. *Nat Rev Micro* **2:** 57-65.

Chen J, Strous M (2013). Denitrification and aerobic respiration, hybrid electron transport chains and co-evolution. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1827:** 136-144.

Chi E, Bartlett DH (1995). An rpoE-like locus controls outer membrane protein synthesis and growth at cold temperatures and high pressures in the deep-sea bacterium Photobacterium sp. strain SS9. *Molecular Microbiology* **17:** 713-726.

Dunn CA, O'Handley SF, Frick DN, Bessman MJ (1999). Studies on the ADP-ribose Pyrophosphatase Subfamily of the Nudix Hydrolases and Tentative Identification of trgB, a Gene Associated with Tellurite Resistance. *Journal of Biological Chemistry* **274:** 32318-32324.

Edwards KJ, Becker K, Colwell F (2012). The Deep, Dark Energy Biosphere: Intraterrestrial Life on Earth. *Annual Review of Earth and Planetary Sciences* **40:** 551-568.

Elshahed MS, Najar FZ, Aycock M, Qu C, Roe BA, Krumholz LR (2005). Metagenomic Analysis of the Microbial Community at Zodletone Spring (Oklahoma): Insights into the Genome of a Member of the Novel Candidate Division OD1. *Applied and Environmental Microbiology* **71:** 7598-7602.

Feder ME, Hofmann GE (1999). Heat-shock proteins, molecular chaperones, and the stress response: Evolutionary and Ecological Physiology. *Annual Review of Physiology* **61:** 243-282.

Fujioka K, Okino K, Kanamatsu T, Ohara Y (2002). Morphology and origin of the Challenger Deep in the Southern Mariana Trench. *Geophysical Research Letters* **29:** 10-11-10-14.

Gabelli SB, Bianchet MA, Bessman MJ, Amzel LM (2001). The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. *Nat Struct Mol Biol* **8:** 467-472.

Gardner JG, Keating DH (2010). Requirement of the Type II Secretion System for Utilization of Cellulosic Substrates by Cellvibrio japonicus. *Applied and Environmental Microbiology* **76:** 5079-5087.

Gihring TM, Zhang G, Brandt CC, Brooks SC, Campbell JH, Carroll S, Criddle, Craig S, Green, SJ, Jardine, P, Kostka, JE, Lowe, K, Mehlhorn, TL, Overholt, W, Watson, DB, Yang, Z, Wu, W, Schadt, CW (2011). A Limited Microbial Consortium Is Responsible for Extended Bioreduction of Uranium in a Contaminated Aquifer. *Applied and Environmental Microbiology* **77:** 5955-5965.

Glud RN, Wenzhofer F, Middelboe M, Oguri K, Turnewitsch R, Canfield DE, Kitazato, H (2013). High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature Geosci* **6:** 284-288.

Harris JK, Kelley ST, Pace NR (2004). New Perspective on Uncultured Bacterial Phylogenetic Division OP11. *Applied and Environmental Microbiology* **70:** 845-849.

Ho TD, Ellermeier CD (2011). PrsW Is Required for Colonization, Resistance to Antimicrobial Peptides, and Expression of Extracytoplasmic Function σ Factors in Clostridium difficile. *Infection and Immunity* **79:** 3229-3238.

Ivancic T, Jamnik P, Stopar D (2013). Cold shock CspA and CspB protein production during periodic temperature cycling in Escherichia coli. *BMC research notes* **6:** 248.

Jones AC, Monroe EA, Podell S, Hess WR, Klages S, Esquenazi E, Niessen, S., Hoover, H, Rothmann, M, Lasken, RS,Yates, JR, Reinhardt, R, Kube, M, Burkart, MD, Allen, EE, Dorrestein, PC, Gerwick, WH, Gerwick, Lena (2011). Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium Lyngbya majuscula.

Proceedings of the National Academy of Sciences **108:** 8815-8820.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas, Brian C, Banfield, JF (2013). Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio* **4**.

Kobayashi H, Hatada Y, Tsubouchi T, Nagahama T, Takami H (2012). The Hadal Amphipod <italic>Hirondellea gigas</italic> Possessing a Unique Cellulase for Digesting Wooden Debris Buried in the Deepest Seafloor. *PLoS ONE* **7:** e42727.

Korotkov KV, Sandkvist M, Hol WGJ (2012). The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat Rev Micro* **10:** 336-351.

Lairson LL, Henrissat B, Davies GJ, Withers SG (2008). Glycosyltransferases: Structures, Functions, and Mechanisms. *Annual Review of Biochemistry* **77:** 521-555.

Lauro FM, Stratton TK, Chastain RA, Ferriera S, Johnson J, Goldberg SM Yayanos, AA, Bartlett, DH (2013). Complete Genome Sequence of the Deep-Sea Bacterium Psychromonas Strain CNPT3. *Genome Announc* **1**.

Lee PT, Hsu AY, Ha HT, Clarke CF (1997). A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the Escherichia coli ubiE gene. *Journal of Bacteriology* **179:** 1748-1754.

Lindquist S (1986). The Heat-Shock Response. *Annual Review of Biochemistry* **55:** 1151-1191.

Löffler FE, Yan J, Ritalahti KM, Adrian L, Edwards EA, Konstantinidis KT, Müller JA, Fullerton H, Zinder SH,Spormann AM (2013). Dehalococcoides mccartyi gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, Dehalococcoidia classis nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi. *International Journal of Systematic and Evolutionary Microbiology* **63:** 625-635.

Markowitz V, Chen I, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner, A Huang, JH, Woyke, T, Huntemann, M, Anderson, I, Billis, K, Varghese, N, Mavromatis, K, Pati, A, Ivanova, NN, Kyrpides, NC (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42:** D560-D567.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi, N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides

genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences* **110:** E2390-E2399.

Melville S, Craig L (2013). Type IV Pili in Gram-Positive Bacteria. *Microbiology and Molecular Biology Reviews* **77:** 323-341.

Merrick MJ (1993). In a class of its own — the RNA polymerase sigma factor σ;54 (σN). *Molecular Microbiology* **10:** 903-909.

Nagata T, Tamburini C, Arístegui J, Baltar F, Bochdansky AB, Fonda-Umani S, Fukuda H, Gogou A, Hansell DA, Hansman RL, Herndl GJ, Panagiotopoulos C, Reinthaler T, Sohrin R, Verdugo P, Yamada N, Yamashita Y, Yokokawa T, Bartlett DH. (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. *Deep Sea Research Part II: Topical Studies in Oceanography* **57:** 1519-1536.

Nakanishi M, Hashimoto J (2011). A precise bathymetric map of the world's deepest seafloor, Challenger Deep in the Mariana Trench. *Marine Geophysical Research* **32:** 455-463.

Nelson SW, Binkowski DJ, Honzatko RB, Fromm HJ (2004). Mechanism of Action of Escherichia coli Phosphoribosylaminoimidazolesuccinocarboxamide Synthetase†. *Biochemistry* **44:** 766-774.

Nunoura T, Takaki Y, Kazama H, Hirai M, Ashi J, Imachi H, Takai, K (2012). Microbial Diversity in Deep-sea Methane Seep Sediments Presented by SSU rRNA Gene Tag Sequencing. *Microbes and Environments* **27:** 382-390.

Peura S, Eiler A, Bertilsson S, Nykanen H, Tiirola M, Jones RI (2012). Distinct and diverse anaerobic bacterial communities in boreal lakes dominated by candidate division OD1. *ISME J* **6:** 1640-1652.

Phadtare S, Inouye M (1999). Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of Escherichia coli. *Molecular Microbiology* **33:** 1004-1014.

Podell S, Gaasterland T (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology* **8**.

Porankiewicz J, Schelin J, Clarke AK (1998). The ATP□dependent Clp protease is essential for acclimation to UV□B and low temperature in the cyanobacterium Synechococcus. *Molecular microbiology* **29:** 275-283.

Price M, Dehal P, Arkin A (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26:**

1641-1650.

Pruesse E, Peplies J, Glockner F (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28:** 1823-1829.

Raina S, Missiakas D, Georgopoulos C (1995). The RpoE gene enconding the sigma (E) (sigma(24)) heat-shock sigma-factor of escherichia coli. *Embo Journal* **14:** 1043-1055.

Redinbo M, Yeates T, Merchant S (1994). Plastocyanin: Structural and functional analysis. *Journal of Bioenergetics and Biomembranes* **26:** 49-66.

Richter K, Haslbeck M, Buchner J (2010). The heat shock response: life on the verge of death. *Molecular cell* **40:** 253-266.

Rinke C, Schwientek P, Sczyrba A, Ivanova N, Anderson I, Cheng J, Darling, A, Malfatti, S, Swan, BK, Gies, EA, Dodsworth, JA, Hedlund, BP, Tsiamis, G, Sievert, SM, Liu, WT, Eisen, JA, Hallam, SJ, Kyrpides, NC, Stepanauskas, R, Rubin, EM, Hugenholtz, P, Woyke, T (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nature protocols* **9:** 1038-1048.

Sandkvist M (2001). Biology of type II secretion. *Molecular Microbiology* **40:** 271-283.

Schauer R, Bienhold C, Ramette A, Harder J (2009). Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME J* **4:** 159-170.

Shi W, Sun H (2002). Type IV Pilus-Dependent Motility and Its Possible Role in Bacterial Pathogenesis. *Infection and Immunity* **70:** 1-4.

Toei M, Gerle C, Nakano M, Tani K, Gyobu N, Tamakoshi M, Sone N, Yoshida M, Fujiyoshi Y, Mitsuoka K, Yokoyama K (2007). Dodecamer rotor ring defines H+/ATP ratio for ATP synthesis of prokaryotic V-ATPase from Thermus thermophilus. *Proceedings of the National Academy of Sciences* **104:** 20256-20261.

Tomoyasu T, Gamer J, Bukau B, Kanemori M, Mori H, Rutman A, Oppenheim, AB, Yura, T, Yamanaka, K, Niki, H (1995). Escherichia coli FtsH is a membrane-bound, ATP-dependent protease which degrades the heat-shock transcription factor sigma 32. *The EMBO journal* **14:** 2551.

Torsvik V, Øvreås L, Thingstad TF (2002). Prokaryotic Diversity--Magnitude, Dynamics, and Controlling Factors. *Science* **296:** 1064-1066.

Wade JT, Roa DC, Grainger DC, Hurd D, Busby SJW, Struhl K, Nudler, E (2006). Extensive functional overlap between [sigma] factors in Escherichia coli. *Nat Struct Mol Biol* **13:** 806-814.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, Brochier-Armanet C, Chain PSG, Chan PP, Gollabgir A, Hemp J, Hügler M, Karr EA, Könneke M, Shin M, Lawton TJ, Lowe T, Martens-Habbena W, Sayavedra-Soto LA, Lang D, Sievert SM, Rosenzweig AC, Manning G, Stahl DA (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proceedings of the National Academy of Sciences* **107:** 8818-8823.

Weisburg W, Barns S, Pelletier D, Lane D (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology* **173:** 697-703.

Whitman W, Coleman D, Wiebe W (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95:** 6578-6583.

Wrighton K, Castelle C, Wilkins M, Hug L, Sharon I, Thomas B, Handley KM, Mullin SW, Nicora CD, Singh A, Lipton MS, Long PE, Williams KH, Banfield JF (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *Isme Journal* **8:** 1452-1463.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield, JF (2012). Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* **337:** 1661-1665.

Wösten MMSM (1998). Eubacterial sigma-factors. *FEMS Microbiology Reviews* **22:** 127-150.

Yoshida M, Muneyuki E, Hisabori T (2001). ATP synthase—a marvellous rotary engine of the cell. *Nature Reviews Molecular Cell Biology* **2:** 669-677.

Yura T, Nagai H, Mori H (1993). Regulation of the Heat-Shock Response in Bacteria. *Annual Review of Microbiology* **47:** 321-350.

Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A, Ramette A (2011). Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLoS ONE* **6:** e24570.

Chapter 4

Genomic characterization of Marinimicrobia (Marine Group A, SAR406) single cell genomes from the Challenger Deep

ABSTRACT

Little is known about the diversity and metabolic capabilities of novel microorganisms from ultra deep ocean environments despite their potential to reveal important information about biogeochemical cycling at depth. The candidate phylum Marinimicrobia (Marine Group A, SAR406) is abundant in the deep-ocean microbial communities, but information about the metabolism of this microorganism remains minimal. Here, we report the analysis of six Marinimicrobia single amplified genomes (SAGs) from sediment obtained from the deepest ocean depth, the Challenger Deep within the Mariana Trench, recovered as part of the Deepsea Challenge Expedition. Phylogenetic and metabolic characterization revealed two distinct Marinimicrobia clades not associated with previously described Marinimicrobia phylogenetic groups, but mostly associated with sequences obtained from deep-sea environments, particularly sediments. Novel metabolic potential based on comparative genomics suggest that these hadal microorganisms take advantage of compounds such as carbon monoxide and hydrogen sulfide to supplement their energy requirements. Genes associated with osmotic and oxidative stress regulationwere also found to be more abundant in the hadal Marinimicrobia. Horizontally transferred genes associated with archaea and eukarya were also found in the Marinimicrobia genomes. Illumina-tag sequencing of bottom water samples collected in additional regions of the Challenger Deep and in the Sirena Deep reinforce the proposition that Marinimicrobia are abundant in Mariana Trench hadal ecosystems.

INTRODUCTION

Estimates of biodiversity among microbial communities within the deep ocean has changed dramatically within the past few decades (Pace, 1997; Sogin *et al*, 2006). Tag pyrosequencing of the V6 region of the 16S rRNA gene identified 'unexpectedly' high *Bacteria* and *Archaea* phylogenetic and functional diversity from deep-sea habitats (Sogin *et al,* 2006; Huber *et al,* 2007). Analyses of the relationship between diversity and ocean depth indicates that diversity increases with water column depth at the phylum/class (Brown *et al*, 2009; Quince *et al*, 2008), and there is also great diversity within deep-sea sediments (reviewed in Orcutt *et al*, 2011). Different groups of organisms are responsible for the cascade of biogeochemical transformations occurring in sediments including some highly diverse groups that are as yet only known as candidate phyla (Schauer *et al*, 2009; Nunoura *et al*, 2012). Examining the deduced metabolic capabilities of novel organisms is one approach to provide a better understanding of nutrient cycling in the ocean (Pedrós-Alió, 2006; Pedrós-Alió, 2011), including within deep and ultradeep seawater and sediments.

The novel candidate phylum (CP) Marinimicrobia was first described as Marine Group A based on three 16S rRNA gene clones recovered from samples collected in the Pacific Ocean off of the San Diego coast and within the Sargasso Sea (Fuhrman *et al*, 1993). A few years later, similar sequences to those described by Fuhrman and colleagues were recovered and described as the SAR406 lineage (Gordon & Giovannoni, 1996). Gordon and Giovannoni proposed that their SAR406-lineage was related to *Chlorobium* and *Fibrobacter* species based on 16S rRNA gene clone libraries prepared from samples collected from two oceans; an 80 m depth sample collected in the western Sargasso Sea at the Bermuda Atlantic Time Series Station and a 120 m depth sample

collected at a site in the Pacific Ocean 70 km from the Oregon coast. Stringent 16 rRNA-based phylogenetic analyses failed to assign the now Marinimicrobia (named by Rinke *et al*, 2013) group to any of the major bacterial phyla. Subsequent phylogenetic analyses placed it closest to the phylum *Caldithrix*, named after a genus of anaerobic, mixotrophic, thermophiles obtained from a hydrothermal vent chimney in the Mid-Atlantic Ridge (Miroshnichenko *et al*, 2003; Rappe and Giovannoni, 2003). Since then phylotypes within the Marinimicrobia cluster have been found in additional environments including deep-sea sediments and oxygen minimum zones (Bowman and Mccuaig, 2003; Fuchs *et al*, 2005; Crump *et al*, 2007; Kato *et al*, 2009). However, the physiological characteristics of the members of these bacteria remain poorly documented due in part to the lack of any cultivated species.

Within the deep ocean, members of the Marinimicrobia CP have been reported from 1000 m to ~4000 m depth in the North Atlantic (Gallagher *et al*, 2004). Phylogenetic analyses have also indicated the presence of Marinimicrobia in the Puerto Rico Trench at 6 km depth (Eloe *et al*, 2011b). Indeed, among the free-living microbial cells the most dominant non-proteobacteria were classified as Marinimicrobia. The presence of high levels of Marinimicrobia in deep-sea environments presents a compelling argument for studying its metabolic properties (Eloe *et al*, 2011b). Metagenome studies suggest that deep-sea microbes possess expanded metabolic potential when compared to cells derived from surface environments (Eloe *et al*, 2011a; Smedile *et al*, 2013). These differences reflect the unique environmental characteristics found in the deep and ultradeep ocean, which includes the lack of sunlight, low temperature, reduced and recalcitrant organic matter and high-hydrostatic pressure.

In an effort to characterize Marinimicrobia, a few phylogenetic and genomic studies have been conducted, particularly in oxygen minimum zones (Allers *et al*, 2013; Wright *et al*, 2013). The most recent comprehensive 16S rRNA gene analysis of marinimicrobial clones, pyro-sequencing and CARD-FISH from the oxygen minimum zone (OMZ) in Northeast subarctic Pacific Ocean (NESAP) revealed that the Marinimicrobia CP is divided into ten subgroups (Allers *et al*, 2013). Recently, Wright and colleagues have analyzed NESAP fosmid clones containing Marinimicrobia sequences and have identified genes involved in fatty acid synthesis, carbon fixation, iron oxidation, the pentose phosphate pathway and genes involved in oxidative stress (Wright *et al*, 2013). Sulfur metabolism genes were also found, in particular, polysulfide reductase genes. These results indicate that cells belonging to the Marinimicrobia CP have the potential to use sulfur compounds as energy sources via respiration of polysulfide to hydrogen sulfide or by dissimilatory oxidation of hydrogen sulfide (Wright *et al*, 2013).

Rinke and colleagues recovered the most substantive genomic data to date for the Marinimicrobia. This was accomplished by sequencing 17 single cell Marinimicrobia genomes from different environments (Gulf of Mexico, terephthalate degrading reactor, sites within the Hawaii Ocean Time-series and South Atlantic Tropical Gyre, and the Etoliko Lagoon, Rinke *et al,* 2013). They proposed the candidate phylum name Marinimicrobia to encompass all SAR406/Marine Group A microbes, and further that they should be placed within the Fibrobacteres–Chlorobi–Bacteroidetes (FCB) super phylum. The FCB super phylum groups a number of diverse organisms, such as the phylum of Bacteroidetes composed by a large number of metabolically diverse bacteria

and the phylum Fibrobacteres, composed mostly of rumen bacteria, based on molecular and phylogenetic signatures (Gupta, 2004).

Among the reported genome sequence deduced properties of the Marinimicrobia single cells are the possession of Ni,Fe hydrogenases and electron transport chain components such as quinol-cytochrome oxidoreductase, cytochrome/quinol oxidase-aerobic, NADH:quinone oxidoreductase, flavoprotein-quinone oxidoreductase and succinate/fumarate:quinone oxidoreductase. These results point towards an aerobic or facultatively anaerobic life style and the possibility of H_2 utilization for energy acquisition.

Here we present a description of the genome properties of six Marinimicrobia single cells derived from surficial sediments within the Challenger Deep. Phylogenetic and comparative genomic analyses are presented along with Illumina-tag sequence data indicating that *Marinimicrobia* are abundant within the Challenger Deep ecosystem. Our results suggest that *Marinimicrobia* cells from the Challenger Deep are phylogenetically distinct to previously described Marinimicrobia subgroups and more closely related to other deep-sea sediment environmental sequences. Metabolic inferences suggest that they are capable of using diverse electron donors and acceptors.

MATERIALS AND METHODS

Collection and sorting

Sediments were collected at 10,908 m depth using a push-core apparatus controlled by a hydraulic arm within in the manned submersible Deepsea Challenger. Sampling occurred during the Deepsea Challenge Expedition on March 26, 2012, in the

east pond of the Challener Deep at 142.59° E, 11.37° N. Recovered sediment was placed in glycerol/TE buffer (Rinke *et al*, 2014) and first stored in liquid nitrogen at -196ºC and later in an ultralow freezer at –80ºC prior to single cell sorting. Samples were transferred to the J. Craig Venter Institute (JCVI) for sorting. The sediment sample was gently vortexed and allowed to settle briefly before filtering through a 35μm mesh (BD Biosciences, San Jose, CA, USA) to avoid larger sediment particles. Cells were stained with 10x SYBR Green I nucleic acid stain (Invitrogen, Carlsbad, CA, USA). Single cells were sorted using a cooled FACS-Aria II flow cytometer and microtiter plates were stored at –80ºC until further processed.

Genome amplification and sequencing

DNA was amplified using a custom BioCel robotic system (Agilent Technologies, Santa Clara, CA) as described by McLean *et al* (2013). Genomic material in the sorted microbial cells was amplified by multiple displacement amplification (MDA) in a 384-well format using a GE GenomiPhi kit (GE Healthcare, Waukesha, WI, USA). 16S rRNA genes were PCR amplified from diluted MDA products using universal bacterial primers 27F and 1492R (Weisburg *et al*, 1991) as follows: 94°C for 3 min, 35 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 90 s, and 72°C for 10 min. PCR products were treated with exonuclease I and shrimp alkaline phosphatase (Thermo Fisher Scientific Inc., Waltham, MA, USA) and sent for Sanger sequencing at the Joint Technology Center (JTC, J. Craig Venter Institute, Rockville, MD, USA). 16S rRNA gene trace files were analyzed and trimmed with the CLC Workbench software program (CLC Bio, Cambridge, MA, USA). Chromatogram quality was assessed manually, and MDAs with both forward and reverse

sequencing primer reads of poor quality were excluded from further analysis. Resulting sequences were evaluated for evidence of microbial DNA contamination associated to MDA reagents, and any samples judged to be contaminated were removed from consideration for whole genome sequencing. Sequences were then compared to the NCBI nr/nt database using BLASTN (Altschul *et al*, 1990) for phylogenetic assignment. DNA recovered from 76 cells was prepared for Illumina sequencing. Libraries were prepared using the multiple barcode technology of the Nextera™ DNA Sample Prep Kit (Illumina, San Diego, CA, USA) and sent to JCT for sequencing. After sequencing, samples were de-multiplexed to separate barcoded sequences for each corresponding single cell genome.

Assembly, annotation and genome completion

Sequenced genomes were processed using Nesoni (www.vicbioinformatics.com/software.nesoni.shtml) and subsequently assembled using the SPAdes 3 assembler (Bankevich *et al*, 2012). Assembled genomes were annotated by IMG-ER (https://img.jgi.doe.gov/cgi-bin/er/main.cgi, Markowitz *et al*, 2014) for complete genome annotation.

16S rRNA gene sequences recovered from each SAG were analyzed by BLASTN against the NCBI nr/nt database (Altschul *et al*, 1990). Sequences with the greatest phylogenetic similarity were extracted from NCBI utilizing the "search and classify" function of the Silva Alignment Service (http://www.arb-silva.de/aligner/; Pruesse *et al*, 2012) and used for phylogenetic reconstruction, along with sequences previously described belonging to the Marinimicrobia from previous publications (Rinke *et al*, 2013;

Allers *et al*, 2012). All sequences extracted from NCBI were also annotated with regard to their associated environmental source, and when it was available, seawater depth. Sequences were aligned with the SINA aligner (http://www.arb-silva.de/aligner/, Pruesse *et al*, 2012) and maximum-likelihood tree were created using FastTree (Price *et al*, 2009). Genome-encoded protein predictions were obtained from IMG-ER and classified phylogenomically using DarkHorse software, version 1.4 (http://darkhorse.ucsd.edu/, Podell and Gaasterland, 2007). DarkHorse was used to predict horizontally transferred genes by assigning a probability that a given protein belongs to the genome being investigated. DarkHorse results were also used to identify potential contaminating sequences among SAG contigs, based on whether or not taxonomic lineages associated with predicted proteins on each assembled contig were similar to or different from the rest of the contigs (Jones *et al*, 2011). Estimated genome completeness for each SAG was calculated as previously described by Rinke *et al* (2013) by using 139 universal single-copy genes. Functional comparisons were performed using the IMG-ER platform (Markowitz *et al*, 2014).

V6 Illumina-tag sequencing

Seawater was collected using Niskin bottles attached to a lander (Hardy *et al,* 2013) deployed in the Challenger Deep West Deep (11.33564N, 142.20113E; 10,897 m) and Middle Deep (11.36902N, 142.43294E; 10,918 m) (Fujioka *et al*, 2002), the Sirena Deep (12.03924 N, 144.34868E; 10,677 m) and the Ulithi Atoll region as a control site (10.00645N, 139.74602E; 761 m). Samples were filtered in series using a peristaltic pump through a 3 μm Isopore filter, followed by a 0.22 μm Sterivex filter, followed by a

0.1 μm Supor filter. Filters were stored in sucrose-Tris lysis buffer at -20C until further processing. DNA was isolated by phenol-chloroform extraction. Briefly, sequencing and curation was done using the VAMPS sequences analysis platform (http://vamps.mbl.edu/index.php; Huse *et al*, 2014). Recovered sequences were process to trim primers and remove low quality reads (Huse *et al*, 2007), then analyzed using the MOTHUR community analysis package (Schloss *et al*, 2009). Within MOTHUR, sequences were aligned to and subsequently classified using the latest greengenes database (DeSantis *et al*, 2006). Classification was done using a 80% similarity cutoff to the greengenes reference database and poorly aligned sequences and chimeras were removed. Classified sequences were clustered at minimum 97% identity.

RESULTS AND DISCUSSION

Genomic properties

From sediment suspension 3520 cells were sorted onto 384 well plates and two plates were amplified by MDA. After PCR of the 16S rRNA gene, 407 of them contained amplicons. Of those, 12 were identified as belonging to the Marinimicrobia (3.7%). Among all samples subjected to genome amplification by MDA, 76 were selected for whole genome sequencing, and of those, six of them were derived from cells within the Marinimicrobia. This affiliation was corroborated based on comparisons of sequenced 16S rRNA genes to the NCBI nt/nr database. For simplicity purposes the SAGs are termed SAR406-CHDEXX. SAR406 refers to the original terminology denoting the phylum, CHDE refers to the CHallenger DEep, and XX refers to numbers

identifying the genomes from 1 – 6. The amount of genome sequence recovered for the six SAGs ranged from 0.9Mbp to 1.9Mbp, with genome completeness extending from 37 to 75 percent, gene counts from 1126 (42% complete) to 2243 (75% complete), and tRNA counts from 15 (42% complete) to 38 (75% complete). The GC content was 41% for SAR406-CHDE6, 42% for four of the SAGs (SAR406-CHDE1, 2, 4, and 5), and 51% for SAR406-CHDE3 (Table 4.1).

The SAR406-CHDE SAGs were compared to the Marinimicrobia genomes reported by Rinke and colleagues (Rinke *et al*, 2013). The percent of the CHDE genomes identified as coding for protein or RNA varied from 88 to 90 percent, which is less than that of all of the 18 Marinimicrobia genomes reported by Rinke and colleagues (2013), indicating that the SAR046-CHDE genomes possess more non-coding/intergenic regions. The number of transposases found in the SAR406-CHDE SAGs are also greater than those found in their comparison Marinimicrobia genomes (Table 4.1). It is also noteworthy that one of the single cell genomes, SAR406-CHDE1, includes a clustered regularly interspaced short palindromic repeat (CRISPR). The CRISPR/Cas gene system is a microbial immunity mechanism that function as a two part process: the immunization process and the immunity process. After entry of exogenous DNA from viruses or plasmids, a Cas complex recognizes the foreign DNA and integrates a novel repeat-spacer unit at the leader end of the CRISPR locus. Later in the immunity process the CRISPR repeat-spacer array is transcribed into a pre-crRNA that is processed into mature crRNAs, which are subsequently used as a guide by a Cas complex to interfere with the corresponding invading nucleic acid (Barrangou and Horvath, 2011). CRISPRs along with Cas genes are considered an adaptive microbial immune response, which provides

acquired immunity against viruses and plasmids (Horvath and Barrangou, 2010). Thus the genomes of the CHDE SAGs contain a greater fraction of noncoding DNA, transposable elements, and phage-related sequences (see also section on horizontal gene transfer below). All properties shared with other deep-sea microbes (Lauro and Bartlett, 2008).

Phylogenetic comparisons

All 12 Marinimicrobia genomes originally recovered from the MDA reactions were used to generate a 16S rRNA gene phylogenetic tree along with reference sequences from reported studies of Marinimicrobia diversity (Allers *et al*, 2013; Rinke *et al*, 2013) (Figure 4.1). All but one of the 16S rRNA genes analyzed clustered together in a clade basal to the root of the CP, indicating that their origins are the most ancient yet described for any members of the Marinimicrobia. When compared against the NCBI nt/nr database by BLASTN all but one of the 16S rRNA sequences were most closely related to sequences obtained from deep-sea sediments or water samples from two sites, the southern edge of the South Pacific Gyre from 5076 m to 5306 m depth (Durbin and Teske, 2010, 2011) and sediments from the Angola Basin in the South-Atlantic Ocean, at water depths ranging from 5032 to 5649 m (Schauer *et al*, 2009). The remaining 16S rRNA sequence, from SAR406-CHDE3, falls within a clade of undescribed environmental sequences. The sequence that was most closely related to SAR406-CHDE3 contained only 92% similarity, and was recovered from subseafloor sediments of the South China Sea at a water depth 3,697 m (Wang *et al*, 2010). All sequences fall within clades outside of those previously recognized within the Marinimicrobia group

(Allers *et al*, 2013), suggesting that these genomes represent two new subgroups within the Marinimicrobia that appear to be biogeographically limited to deep-sea sediments.

Metabolic profiles

For greater simplicity the two distinct phylogenetic groups of Marinimicrobia uncovered within the Challenger Deep are referred to as clades A (SAR406-CHDE1, 2, 4, 5 and 6) and B (SAR406-CHDE3). The presence of genes for the transport and metabolism of carbohydrates via glycolysis, the tricarboxylic citrate acid (TCA) cycle and the non-oxidative branch of the pentose phosphate pathway, points towards a general heterotrophic lifestyle among all the SAR406-CHDE cells. However, it is important to note that two of the genomes in group A (SAR406-CHDE4 and 6) encode an 2-oxoacid:acceptor oxidoreductase, one of the key enzymes involved in the reductive tricarboxylic acid (rTCA) cycle used for carbon fixation. Other enzymes required for various modes of carbon fixation in microorganisms are also encoded in the genomes of group A; acetyl/propionyl-CoA carboxylase, alpha subunit involved in the 3-hydroxypropionate/4-hydroxybutyrate cycle and phosphoenolpyruvate carboxylase, type 1 (EC 4.1.1.31) involved in the dicarboxylate/4-hydroxybutyrate cycle. However, these enzymes could also be involved in other metabolic processes within the cell, and due to the incompleteness of the single cell genomes it is not possible to accurately assess the presence or absence of carbon fixation in these SAR406 CHDE genomes. The SAR406-CHDE cells appear to be capable of aerobic as well as anaerobic metabolism. The former is represented by genes whose components function in the respiratory chain or as cytochrome oxidoreductase, the latter by genes associated with pyruvate fermentation to

lactate and alcohol (SAR406-CHDE1, 2 and 4), and amino acid fermentation involving shikimate dehydrogenase (EC:1.1.1.25) (SAR406-CHDE1, 2 and 4).

Interestingly, the clade B genome SAR406-CHDE3 is the only genome that encodes components of a nitrate reductase (NarG nitrate reductase alpha subunit 67% similar to *Geothrix fermentans*, NarH nitrate reductase beta subunit 64% similar to *Geothrix fermentans* and nitrate reductase chaperone NarJ 48 % similar to *Geothrix fermentans*). *Geothrix fermentans* is an, anaerobic bacterium normally be found in aquatic sediments (Coates *et al*, 1999). SAR406-CHDE3 encodes for a respiratory nitrate reductase, which catalyzes the first step in the denitrification pathway proceeding from NO_3 to N_2, a process that is mostly associated with anaerobic metabolism using nitrate as an electron acceptor (Knowles, 1982). SAR406-CHDE3 does not possess any of the other genes associated with denitrification, but its genome is only 37% complete, thereby precluding a complete description of its metabolic potential. On the other hand SAR406-CHDE3 also encodes for components of the cytochrome oxidase involved in aerobic respiration. Taken together these results indicate that the clade B SAG SAR406-CHDE3 is likely to encode the ability to respire both oxygen and nitrate.

Energy acquisition

As is common to many deep-sea bacteria (Reinthaler *et al*, 2010), auxiliary energy yielding pathways are also present in these genomes. It appears that the SAR406-CHDE members of clade A possess genes used to derive energy from the oxidation of carbon monoxide. All CHDE genomes in clade A encode the large, middle and small subunits of the aerobic-type carbon monoxide dehydrogenase (CODH; CoxL, CoxM and CoxS).

When compared to other CODH enzymes, the SAR406-CHDE CODHs are more closely related to putative CODHs and other dehydrogenases in the molybdenum hydroxylase family. Some of the enzymes most similar to the SAR406-CHDE CoxL subunits retrieved from BLASTP are annotated as aldehyde oxidase and xanthine dehydrogenase from *Kosmotoga olearia,* an anaerobic heterotroph capable of hydrocarbon oxidation coupled with sulfate reduction (DiPippo *et al*, 2009) or aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL-like protein are from *Mesotoga prima, an* anaerobic microorganism that utilizes sulfur compounds as electron acceptors (Nesbø *et al*, 2012). BLASTP analyses of the middle subunit SAR406-CHDE CODHs are more conclusive, showing similarity exclusively with annotated carbon monoxide dehydrogenases from *Desulfurococcus kamchatkensis,* an anaerobic heterotrophic hyperthermophilic crenarchaeon isolated from a terrestrial hot spring (Ravin *et al*, 2009). The phylogenetic association of CODH with archaeal organisms suggests that these genes were horizontally transferred (see section on horizontal gene transfer below). CODH genes are also encoded by some of the Marinimicrobia comparison genomes (5 of the 18 encode CoxL and 4 out of 18 encode CoxM and CoxS). Interestingly, all of the comparison genomes that possess CODH were retrieved from the terephthalate degrading reactor sample.

The conversion of CO to CO_2 generates electrons that can be shuttled to the respiratory chain for energy generation and CO_2 that can be used for carbon fixation (Ferry, 1995). In the case of aerobic CO oxidation, carbon fixation is accomplished by the Calvin-Benson cycle (Anand and Satyanayarana, 2012). Because the SAR406 CHDE

SAGs do not contain genes involved in the Calvin-Benson cycle it is most probable that carbon monoxide dehydrogenase is being used solely to generate electrons used in aerobic respiration (King and Weber, 2005).

The SAR406-CHDE genomes in clade A also encode for a number of genes involved in sulfur metabolism including a sulfide-quinone oxidoreductase (EC 1.8.5.4) (SAR406-CHDE1 and 2). This enzyme is responsible for the oxidation of hydrogen sulfide to sulfur, which may be involved in sulfide detoxification and sulfide-dependent respiration. In the latter case the electrons generated in this process are transferred to the electron transport chain for energy conservation. Another method for energy production apparently used by the SAR406-CHDE microbes may come from the conversion of hydrogen sulfide to elemental sulfur or sulfate via sulfide-quinone oxidoreductase (SQR). SQR has been reported in a diverse range of microorganisms (Theissen *et al*, 2003), including a species closely related to Marinimicrobia, *Chlorobium limicola* (Peschek *et al*, 1999). Studies conducted in the hyperthermophilic and chemolithoautotrophic bacterium *Aquifex aeolicus* have led to the conclusion that SQR is involved in sulfide-dependent respiration (Nubel *et al*, 2000). Biological sulfide oxidation may be an important process for the global circulation of sulfur in various oxic–anoxic interface environments (Griesbeck *et al*, 2000). This enzyme was not found in any of the comparison Marinimicrobia genomes.

Genomic comparisons

The SAR406-CHDE genomes were compared to single cell genomes available in the IMG-ER platform previously described by Rinke and colleagues (Rinke *et al*, 2013).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Clusters of Orthologous Groups (COGs) and enzyme (EC) presence or absence was assessed with a special focus on the genes unique to the SAR406-CHDE genomes. Among the characteristics shared between all Marinimicrobia genomes are abilities for utilizing carbohydrates for energy production via glycolysis, the TCA cycle, and the non-oxidative branch of the pentose phosphate pathway. The SAR406-CHDE genomes share genes involved with oxidative phosphorylation with some, but not all of the comparison Marinimicrobia genomes. Those Marinimicrobia comparison genomes that lack genes for oxidative phosphorylation may do so as a result of the incomplete status of their sequences, or because they are derived from cells that rely on fermentation for all their energy needs. Similar concerns about the incompleteness of genome sequences applies to all cases where categories of genes are absent, this is also the case for peptidoglycan and lipopolysaccharide synthesis genes. SAR406-CHDE genome (from clade A and B) as well as comparison Marinimicrobia genomes encode for gene in the peptidoglycan and lipopolysaccharide biosynthetic pathways. These suggest that the Marinimicrobia genomes analyzed here can be defined as Gram negative bacteria.

Some genes were found to be shared among all SAR406-CHDEs and absent in all of the comparison genomes, bolstering the case that they are truly SAR406-CHDE specific. All of the SAR406-CHDEs encode a component of the NitT/TauT family transport system (KO:K02051) that is not present in the other genomes. This system is involved in nitrate/nitrite/cyanate and taurine uptake (Saier, 2000). Also shared by all of the CHDE SAGs is a two-component system (TCS) within the NtrC family, the nitrogen regulation response regulator NtrX. It is believed to be involved in the regulation of

respiratory gene expression in bacteria, and most likely responds to signals arising from oxygen limitation. This TCS may be critical for bacterial survival under conditions where oxygen is limiting (Atack *et al*, 2013).

Another enzyme uniquely shared by all of the CHDE cells is cyclic pyranopterin phosphate synthase (EC: 4.1.99.18), which is involved in the biosynthesis of molybdenum cofactor (MoCo). The MoCo forms the active site of almost all molybdenum (Mo) containing enzymes. One of the molybdenum-dependent enzymes is the pterin-based enzyme xanthine oxidase (XO), which is involved in purine catabolism as well as cellular responses to senescence and apoptosis (Hillie, 2005). Most enzymes of the XO family are well characterized as purine-oxidizing and/or aldehyde-oxidizing enzymes with broad substrate specificities, but several more specific enzymes, such as carbon monoxide dehydrogenase have also been described (Meyer and Rajagopalan, 1984). A number of other genes involved in MoCo biosynthesis are also present in many of the SAR406-CHDE genomes and absent from the comparison genomes, among them molybdenum cofactor biosynthesis protein B, molybdopterin molybdotransferase (EC:2.10.1.1), molybdopterin synthase catalytic subunit (EC:2.8.1.12) and molybdopterin synthase sulfur carrier subunit. The presence of the biosynthetic machinery for MoCo, bolsters the case that the SAR406-CHDE genomes encode a functional carbon monoxide dehydrogenase.

Osmotic regulation

Genes associated with osmotic regulation, like aquaporin Z (AqpZ), are uniquely found in the SAR406-CHDE genomes, suggesting that osmotic regulation is important

for deep-sea adaptation. For example, an aquaporin Z gene is shared between four of the genomes in clade A (SAR406-CHDE1, 2, 4 and 6). The role of AqpZ in free-living marine microorganisms has not been fully characterized. However, it has been discovered and characterized in Escherichia coli to function as a channel for rapid water efflux across the membrane, helping microorganisms to cope with osmotic downshift (Calamita, 2000). This membrane channel has a role in both short-term and long-term osmotic adaptation based on its ability to transport water in either direction. Osmotic pressure and hydrostatic pressure can have opposing effects on macromolecules (Robinson *et al*, 1995), and deep-ocean metazoans accumulate large amounts of organic osmolytes to cope with high-pressure conditions (Yancey *et al*, 2014) (sometimes referred to as piezolytes; Martin *et al,* 2002). Aquaporins could play a role in high-pressure adaptation by increasing osmotic concentrations in the cells to balance high hydrostatic pressure influences on protein hydration (Figure 4.2). Aquaporin-like genes have been also reported to be present in single cells collected and analyzed from the deepest part of the Atlantic Ocean, the Puerto Rico Trench (Leon Zayas *et al*, in review). These findings reinforce the potential importance of osmotic regulation in deep-sea environments.

Interestingly, there are other genes that are not shared with the comparison genomes that are also seem to be also involved in osmotic regulation, for example genes involved in an osmoprotectant transport system (opuA and opuBD; Kempf and Bermer, 1995) and an osmotically inducible protein OsmC (Park *et al*, 2008). The osmoprotectant transport system for proline/glycine betaine (OpuA, OpuC and OpuD) is known to be expressed in response to increasing osmotic pressure (Kappes *et al*, 1999) and the osmotically inducible protein C (OsmC) is involved in the cellular defense

mechanism against both oxidative stress caused by exposure to hydroperoxides or to elevated osmolarity (Atichartpongkul *et al*, 2001; Lesniak *et al*, 2003). The presence of genes associated with these systems in the SAR406-CHDE SAGs but not their comparison gneomes provides an even more compelling case for the need for osmoregulation in the Challenger Deep ecosystem.

Horizontally transferred genes

In order to better understand the overall genome similarities among all the SAR406-CHDE SAGs, including genes arising from both vertical and horizontal transmission, the predicted proteins encoded by each of the SAGs were compared to the NCBI database by BLAST. The top hit for each protein prediction was extracted and classified based on its taxonomic association. The most abundant top hit for all of the genomes was to *Melioribacter roseus* P3M 2, a facultatively anaerobic thermophilic cellulolytic bacterium from the class *Ignavibacteria* within the phylum Chlorobi (Podosokorskaya *et al*, 2013). This result reinforces the Marinimicrobia association with the FCB super phylum. The second and third most frequent BLAST matches are to *Ignavibacterium album* and *Candidatus Latescibacter anaerobius*, also members of the FCB super phylum. However, when comparing the distribution of the 33 most frequent BLAST matches many are distinguishable from Marinimicrobia phylogenetic position and likely represent genes acquired by horizontal gene transfer. The similarities between the SAR406-CHDE are shown by their relationships in space on a non-metric multidimensional scaling (nMDS) plot (Figure 4.3). Difference between the genomes observed in the ordination plot suggest that most of the SAGs cluster together in a single

cluster (Figuer 4.3, orange), in agreement with their 16S rRNA phylogenetic association. Although this evaluation of the total genome similarity, encompassing both vertically and laterally transferred genes, is limited by the incompleteness of the genome sequences, the results are consistent with variation in genes introduced by horizontal gene transfer among the SAR406-CHDE SAGs. As an example, genomes from clade A exist closely together at the left of the plot while SAR406-CHDE3 is removed from this cluster (Figure 4.3, blue). It is possible that the differences between the CHDE genomes come from actively exchanging genomic material with other cells. The most abundant top BLAST matches not reflecting SAR406 phylogenetic position was to *Clostridium species,* anaerobic, sporeforming gram-positive microorganisms within the phylum *Firmicutes*.

Lateral gene transfer is one of the main mechanisms for microbial genomic diversification and innovation (Nakamura *et al*, 2004; Ochman *et al*, 2000).
A number of sequences that appear to have been acquired by horizontal gene transfer (HGT) that are most closely related to archaeal and eukaryal organisms are also present in the SAR406-CHDE genomes. This includes 240 archaea-like genes and 70 eukarya-like genes (Table 4.2). Analyses of HGT were performed using DarkHorse software, which predicts HGT by assigning a probability score that a given encoded protein belongs to the genome being investigated (Podell and Gaasterland, 2007). When looking at the taxonomic associations of the putative HTG encoded proteins 83% of archaeal associated HTGs are more similar to phylum Euryarchaeota and among those 32% belong to the class Methanomicrobia. The class Methanomicrobia encompasses a diverse group of archaea including psychrophilic, thermophilic, halophytic, methylotrophic and

methanogenic organisms (Pikuta, 2011). In terms of the eukarya, the most abundant phylum is the Viridiplantae, specifically the class chlorophyta, which is a division of green algae (Leliaert et al, 2012). Among the putative HGTs from archaea there is not one gene that is shared among all of the SAR406-CHDEs, but most of these genes are shared among the 5 phylogenetically distinct clade A SAR406-CHDE genomes. Among the shared HGTs are genes for pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96), used in aromatic amino acid degradation (Naponelli *et al*, 2008) and cysteine synthase (EC 2.5.1.47) that participates in cysteine biosynthesis (Kredich and Tomkins, 1996). These genes involved in amino acid metabolism may represent a need to synthesis and recycle molecules that are difficult to acquire due to the depleted nature of bioavailable organic matter in the deep ocean (Aristegui *et al*, 2009).

A gene for cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2.5.1.17), involved in the biosynthesis of cobalamin (vitamin B12), also appears to have been obtained via HGT. Vitamin B12 is an important cofactor present in a number of key metabolic pathways, including the TCA cycle (Raux *et al*, 2000), where it is used by isocitrate dehydrogenase (NADP) (EC 1.1.1.42), a gatekeeper enzyme that controls metabolic flux between the TCA cycle and the glyoxylate cycle.

Among the HTGs that are more closely related to members of the eukarya are genes involved in biosynthesis of amino acid arginine by ornithine carbamoyltransferase. This gene appears to be most closely related to sequences present in green algae (cold adapted Coccomyxa subellipsoidea; Blanc *et al,* 2012).

Also a small number of phage-like genes were recovered from the SAR406-CHDE genomes (Table 4.3). SAR406-DSC5 has the most number of phage genes, which

are associated with phage assembly and appear to have been acquired by a relative of enterobacterial phage M13 within the Inoviridae family. The other SAR406 genomes have one or two phage-like genes, mostly related to relatives of Cronobacter phage CR9 within the Myoviridae family.

Relative abundance of SAR406 in the Mariana Trench by Illumina-tag sequencing

The biodiversity of the microbial community within the Challenger Deep was assessed by V6 Illumina-tag sequencing of filtered bottom seawater samples recovered during the Deepsea Challenge Expedition. Three different sites within the Mariana Trench were collected for analyses utilizing a deep sea sampling lander. These were the West Deep (10,897 m) and Middle Deep (10,918 m) sections of the Challenger Deep, and the Sirena Deep (10,677 m), along with a shallower (761 m) reference site. The significance and potential metabolic importance of the Marinimicrobia CP in the hadal environment of the Mariana Trench, is supported by the fact that tag sequences associated to the Marinimicrobia CP were one of the most abundant operational taxonomic units (OTU) within the Mariana Trench water/sediment interface samples, extending to greater than 9% abundance in one case (Figure 4.4). Indeed, clustering and classification at the phylum level revealed that except for the phylum proteobacteria, the Marinimicrobia CP was the most abundant group in almost all size fractions of each of the deep-sea samples. It was also more abundant in the trench locations relative to the shallow-water reference site. Given its high abundance the inferred metabolic processes deduced from the SAR405-CHDE SAGS, such as the oxidation of carbon monoxide and sulfide, are likely

to represent significant biogeochemical transformations at hadal depths within this trench system.

CONCLUSION

This study represents the most in-depth description of deep-ocean Marinimicrobia to date. For a microbial group that appears to be abundant in many different environments, including in deep (Smedile *et al*, 2013) and ultra-deep (Eloe *et al*, 2011; Tarn *et al*, unpublished) ocean settings, the lack of physiological data for this CP is striking. The results presented here suggest that the CP Marinimicrobia are mostly heterotrophic organisms, although the possibility of mixotrophy is present in some of the SAR406-CHDE cells. Many of the genomes possess both respiratory and fermentative genomic signatures, which leads to the conclusion that Marinimicrobia are facultative anaerobes. Supplementing energy acquisition by the oxidation of carbon monoxide or hydrogen sulfide may be used, and is more prevalent in deep-sea Marinimicrobia than those members from other habitats. Genes associated with adaptation to osmotic pressure fluctuations also appear to be more prevalent in the deep-sea Marinimicrobia genomes examined in this study, perhaps functioning to help counterbalance the effects of extreme hydrostatic pressure. A large number of genes appear to have been acquired from archaea and eukarya.

ACKNOWLEDGEMENTS

We are grateful for the financial support provided by the National Science Foundation (0801973 and 0827051), a National Science Foundation Graduate Research

Fellowship (068775), the National Aeronautics and Space Administration (NNX11AG10G), a National Institutes of Health Marine Biotechnology Training grant (T32GM067550) and a gift from Earthship Productions. We are specially grateful to James Cameron for his contribution to the collection of these samples.

Chapter 4 is a full-length manuscript in preparation for publication: Rosa León Zayas, Logan Peoples, Jonathan Tarn, Sheila Podell, Mark Novotny, Roger S. Lasken and Douglas H. Bartlett. 'Genomic characterization of Marinimicrobia (Marine Group A, SAR406) single cell genomes from the Challenger Deep' with permission from all coauthors

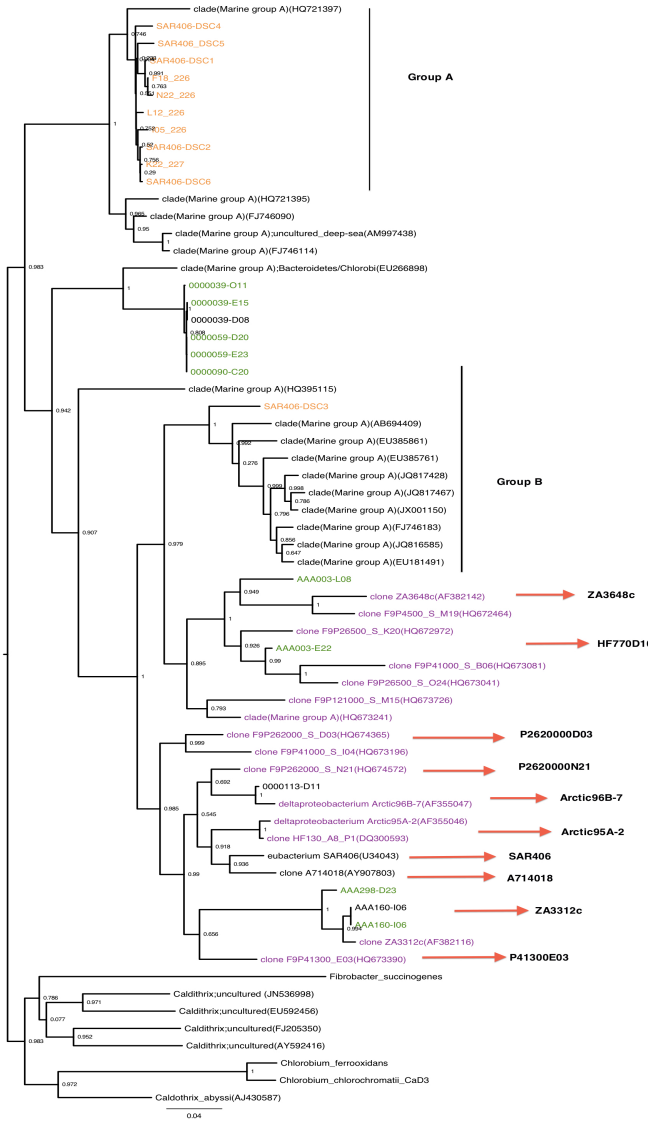

Figure 4.1 Phylogenetic tree of 16S rRNA gene from of SAR406-CHDE SAGs
Rooted maximum likelihood phylogenetic tree of 16S rRNA gene for eleven MDAed Marinimicrobia cells and related uncultured organisms are shown. SAR406-CHDE are highlighted in orange. Sequences highlighted in purple represent previously described Marinimicrobia CP members and 10 suggested phylogenetic subgroups are annotated (Allers *et al*, 2013). Sequences highlighted in green represent other single cell genomes from Gulf of Mexico, Terephthalate degrading reactor, Hawaii Ocean Time-series, Tropical Gyre Atlantic and Etoliko Lagoon (Rinke *et al*, 2013). Two new clades are highlighted, Clade A and Clade B, which include the SAR-CHDE SAGs. Scale bar represents 0.04 changes per position. The displayed confidence values are those that are 50% or lower.

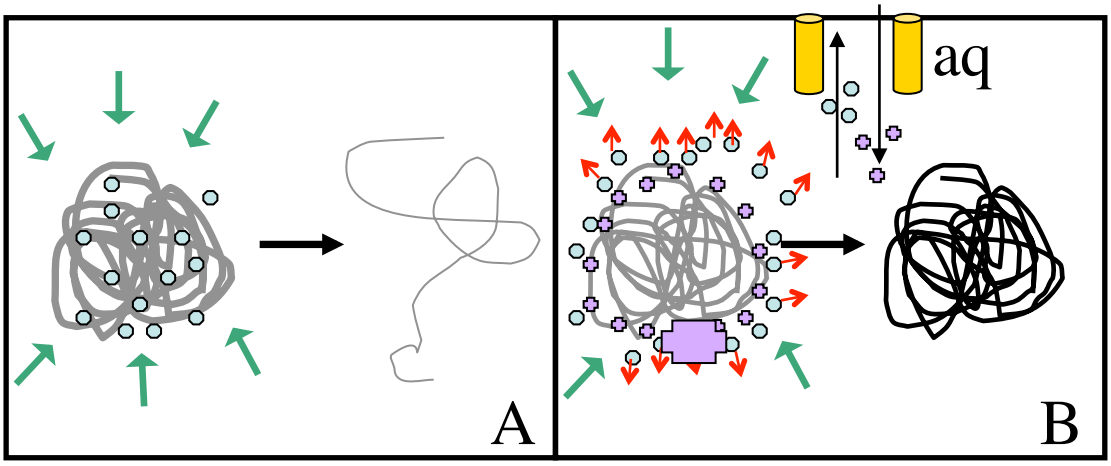

Figure 4.2 Schematic of potential role of aqpZ in pressure adaptation
Diagram representation of hypothetical effect of water efflux, influx of osmolytes and its balancing effect on hydrostatic pressure denaturizing nature. A) Hydrostatic pressure is through to affect protein conformation by destabilizing the water conformation that surround them and forcing water molecules into the proteins empty spaces. B) Aquaporins (aqp) act as a water channel involved in osmoregulation cell through out all domains of life. The increase of osmotic pressure in the cells may counter act the effect of hydrostatic pressure, as the increase in osmolarity induces the release of water molecules from the protein. Aquaporins may also act as a channel for influx of osmolytes, which will also counter act the effect of high hydrostatic pressure. Green arrows represent hydrostatic pressure, blue circles represent water molecules, red arrows represent effect of osmotic pressure, purple pluses represent osmolytes.

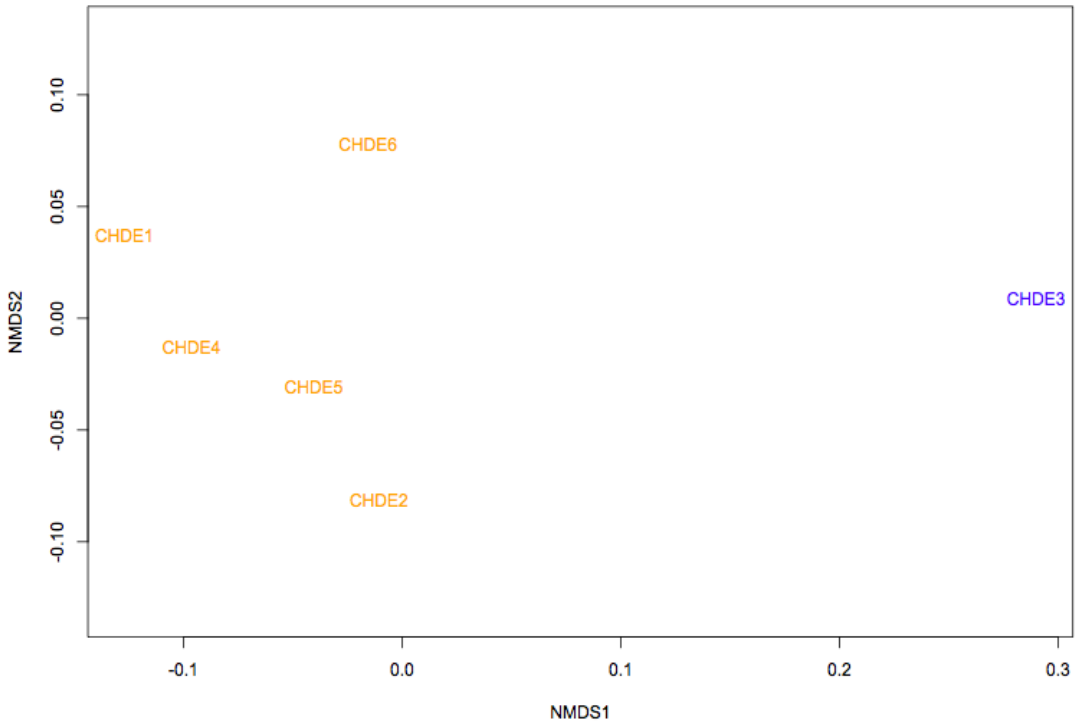

Figure 4.3 Non-Metric Multidimensional Scaling of top species hit SAR406-CHDE genomes Data for the taxonomicassociation of each predicted protein was retrieved from the DarkHorse BLAST analysis. The abundance of each predicted protein top hit microorganism match was calculated and the most abundant organisms (33) were used to assess the similarities among the 6 CHDE genomes. A Non-Metric Multidimensional Scaling (nMDS) ordination was calculated using the R-package *Vegan* a Bray-Curtis algorithm. Top hit values for each genome were normalized for the total number of proteins analyzed by DarkHorse genome. Based on their similarity matrix the data was grouped in two clusters are color-coded based on their relatedness. Two colors were used: blue and orange. Organisms in the orange cluster belong to clades clade A while organisms in the blue belong to clade B. Their distance on the ordination plot mirrors that of their 16S rRNA phylogeny. Stress value is 0.0004, which provided a measure of the fit of the data reported on a range of 0-1. Ordinations with stress higher than 0.3 can't be reliably interpreted; lower stress means the solution fits the data better

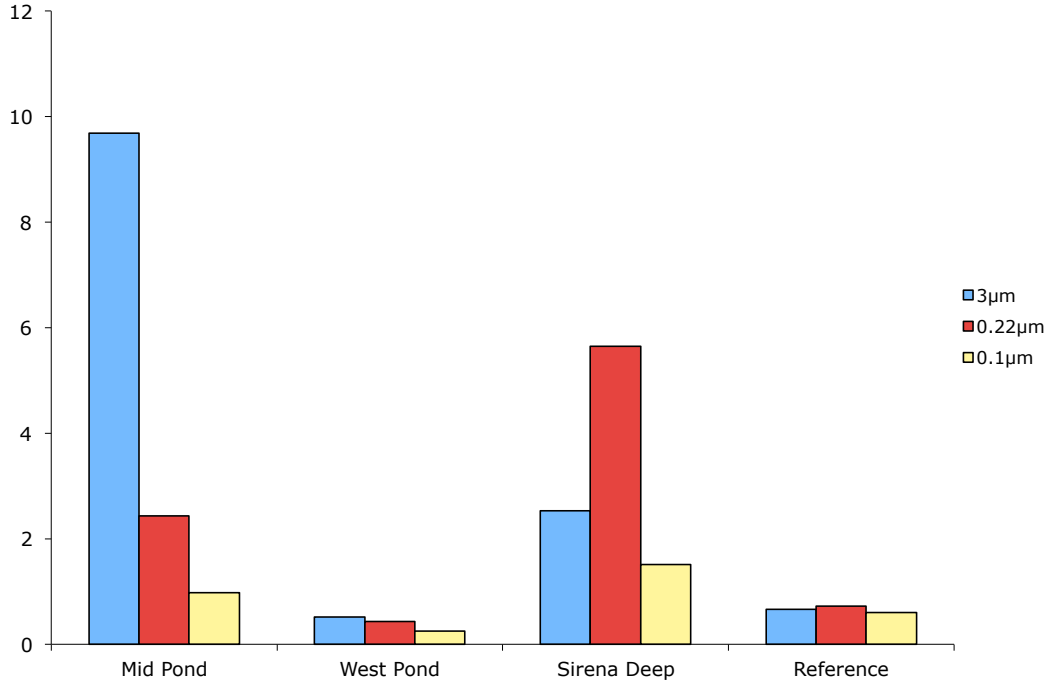

Figure 4.4 Relative abundance of Marinimicrobia among Mariana Trench deep-sea water V6 Illumina-tag sequences
Relative abundance of the Marinimicrobia candidate phylum obtained from each of three different filter fractions in each of the three stations (and control). The relative abundance of the 3μm, 0.2μm and 0.1μm fractions for the four sites are shown in blue, red and cream, respectively. The x-axis displays the four sites and the y-axis displays the percentage of relative abundance within the whole bacterial community.

Table 4.1 Genomic properties of 13 OD1-DSC SAG genomes
The values presented are sequenced genome size, percent completeness, GC percentage, 16S rRNA gene count, CRISPR count, tRNA count, transposase count, percent of coding bases, and horizontally transferred genes from archaeal and eukaryal hosts are displayed for six SAR406-CHDE genomes and comparison Marinimicrobia genomes (Rinke et al, 2013)

Sample Name	Genome Size	% Completeness	GC %	Gene Count	Coding Base Count %	16S rRNA Count	CRISPR Count	tRNA Count	Transposases	HTG (Arch,Euk)
SAR406-CHDE14	1864925	72	42	2020	89.13	1	1	36	7	45,9^
SAR406-CHDE15	1682587	73	42	2143	89	1	0	23	4	39,8^
SAR406-CHDE16	1059145	37	51	1451	88.51	1	0	21	9	21,12^
SAR406-CHDE17	1906748	75	42	2243	88.77	2	0	38	16	54,14^
SAR406-CHDE18	1452644	65	42	1823	90.26	1	0	24	8	47,13^
SAR406-CHDE19	973538	42	41	1126	89.78	1	0	15	9	34,14^
0000039-D08 (Combined_Assembly)	2357096	97	47	2158	93.56	1	1	42	1	26, 1*
0000039-E15 (TAsludge)	493984	36	47	461	94.59	1	0	11	0	0 *
0000039-O11 (TAsludge)	828046	27	47	749	93.31	1	0	16	2	0 *
0000059-D20 (TAbiofilm)	772719	47	48	718	93.24	1	0	18	2	0 *
0000059-E23 (TAbiofilm)	618060	15	47	583	93.78	1	0	8	0	0 *
0000059-L03 (TAbiofilm)	597650	51	48	585	93.01	0	0	5	0	0 *
0000077-B04 (TAbiofilm)	496915	12	47	479	92.97	0	0	7	0	0 *
0000090-C20 (Etoliko)	566695	21	48	511	93.99	1	0	13	0	0 *
AAA003-E22 (Tropical gyre)	888773	51	37	788	92.18	2	2	19	0	0 *
AAA003-L8 (Tropical gyre)	1265631	62	30	1258	95.74	1	2	25	0	0 *
AAA011-A05 (DUSE)	237314	9	48	230	93.84	0	0	0	0	4, 1*
AAA076-M08 (Gulf_of_Mexico)	392252	66	33	440	97.05	0	0	12	0	0 *
AAA160-B08 (Gulf_of_Mexico)	922550	70	33	995	96.34	0	0	26	0	0 *
AAA160-C11 (Gulf_of_Mexico)	824595	86	33	883	96.75	0	0	29	0	0 *
AAA160-I06 (Gulf_of_Mexico)	884929	98	32	1007	96.79	1	0	32	0	0 *
AAA160-I06 (Combined_Assembly)	1119009	100	33	1221	96.8	1	0	36	0	12, 5*
AAA257-N23 (Etoliko)	948616	42	39	912	91.02	0	1	13	4	0 *
AAA298-D23 (Hawaii_Ocean_Time_Serious)	979176	100	31	1055	97.13	1	0	33	0	0 *

^ Recovered from DarkHorse
* Recovered from IMG-ER

Table 4.2 Horizontally transferred genes from archaea or eukarya to the SAR406-CHDE genomes

Genes with best BLAST matches to archaea or eukarya are displayed including the query annotation by IMG, the top matchspecies and the top match gene product function, as assessed by the DarkHorse analysis, for the genes highlighted within the article.

query_annotation	Archaea species	besthit_annotation
SAR406-DSC14		
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96)	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homolo	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2.5.1.17)	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
cysteine synthase (EC 2.5.1.47)	Candidatus Caldiarchaeum subterraneum	cysteine synthase A
SAR406-DSC15		
Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homolo	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2.5.1.17)	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
SAR406-DSC16		
Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family	Pyrobaculum sp. 1860	xanthine dehydrogenase accessory factor
Por secretion system C-terminal sorting domain	uncultured marine crenarchaeote HF4000_APKG2	hypothetical protein
SAR406-DSC17		
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96)	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2.5.1.17)	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
cysteine synthase (EC 2.5.1.47)	Candidatus Caldiarchaeum subterraneum	cysteine synthase A
Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homolo	Desulfurococcus kamchatkensis 1221n	FAD-binding molybdopterin dehydrogenase
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyl transferase 4-like	Methanothermobacter thermautotrophicus CaT2	glycosyltransferase
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
Por secretion system C-terminal sorting domain	uncultured marine crenarchaeote HF4000_APKG2	putative fibronectin type III domain protein
SAR406-DSC18		
Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homolo	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Por secretion system C-terminal sorting domain	Methanosaeta concilii GP6	hypothetical protein
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
SAR406-DSC19		
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96)	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2.5.1.17)	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
Predicted glycosylase	Ferroglobus placidus DSM 10642	glycosidase PH1107-related protein
Predicted glycosyltransferases	Methanobrevibacter ruminantium M1	glycosyl transferase GT2 family

query_annotation	Eukaryote species	besthit_annotation
SAR406-DSC14		
Ornithine carbamoyltransferase	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein
SAR406-DSC16		
Ornithine carbamoyltransferase	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
SAR406-DSC17		
Por secretion system C-terminal sorting domain	Aureococcus anophagefferens	hypothetical protein
SAR406-DSC18		
ornithine carbamoyltransferase (EC 2.1.3.3)	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
SAR406-DSC19		
Por secretion system C-terminal sorting domain	Aureococcus anophagefferens	hypothetical protein
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein

Table 4.3 Phage-like genes found in SAR406-CHDE genomes
Genes with best BLAST matches to viral sequences are displayed in this table including the product name annotation by IMG, the top hit species and the top hit BLAST match assessed by the DarkHorse analysis.

query_annotation	species	lineage	besthit_annotation
SAR406-DSC14			
Uncharacterized small protein	Cronobacter phage CR9	Myoviridae	hypothetical protein
SAR406-DSC15			
Chaperone of endosialidase	Bacillus phage CampHawk	Myoviridae	tailspike
exonuclease, DNA polymerase III, epsilon subunit family	Clostridium phage phiMMP04	Myoviridae	DNA polymerase III alpha subunit
SAR406-DSC17			
Uncharacterized small protein	Cronobacter phage CR9	Myoviridae	hypothetical protein
SAR406-DSC18			
Uncharacterized small protein	Cronobacter phage CR9	Myoviridae	hypothetical protein
Phage major coat protein, Gp8	Enterobacteria phage M13	Inovirus	structural protein
hypothetical protein	Enterobacteria phage M13	Inovirus	small hydrophobic protein
phage/plasmid replication protein, gene II/X family	Enterobacteria phage M13	Inovirus	hypothetical protein
hypothetical protein	Enterobacteria phage M13	Inovirus	phage assembly protein
Type II secretory pathway, component PulD	Enterobacteria phage M13	Inovirus	phage assembly protein
Beta-propeller domains of methanol dehydrogenase type	Enterobacteria phage M13	Inovirus	gene III
Helix-destabilising protein	Enterobacteria phage M13	Inovirus	helix destabilising protein
Zonular occludens toxin (Zot)	Enterobacteria phage M13	Inovirus	phage assembly protein
SAR406-DSC19			
Site-specific recombinase XerD	Pseudomonas phage vB_PaeS_PM(	Siphoviridae	unnamed protein product
Putative viral replication protein	Silurus glanis circovirus	Circovirus	replication-associated protein

Supplementary Material

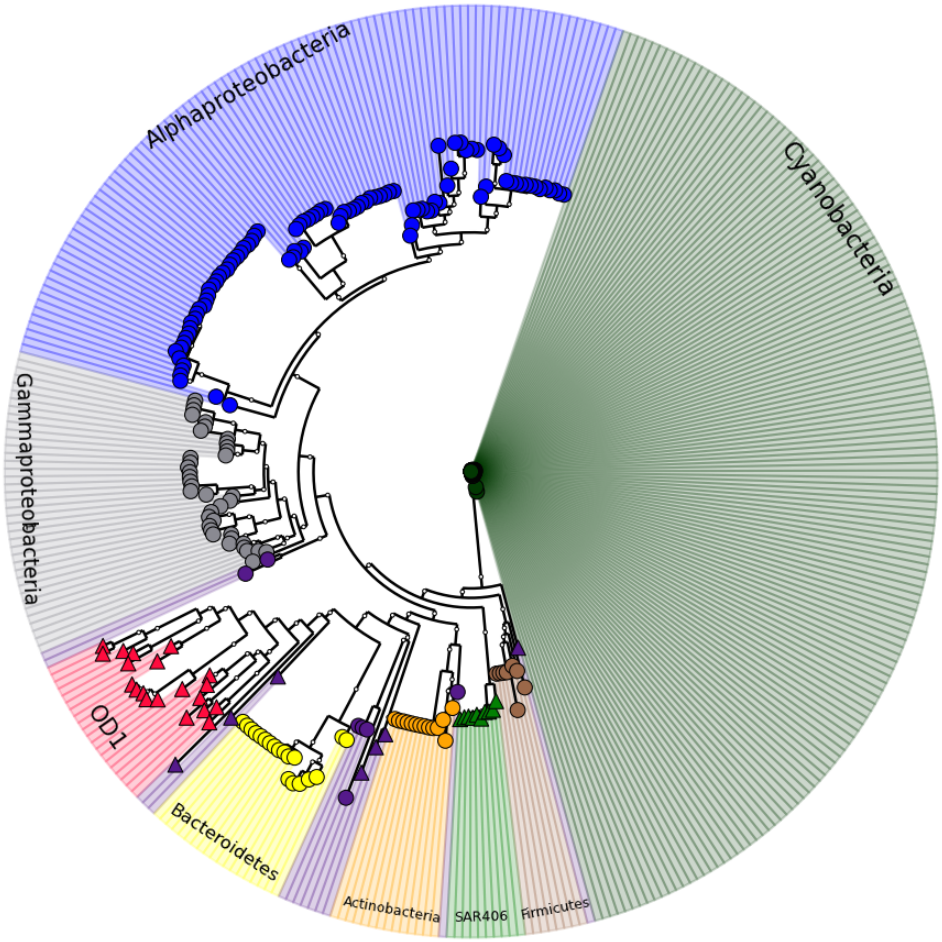

Figure S4.5 Phylogenetic distribution of Challenger Deep MDAs
Tree shows both the phylogenetic distribution of sorted and successfully amplified samples and their relative abundances. The major players are annotated and colored differently. Of a total of 371, fourteen phyla were represented: *Proteobacteria, Cyanobacteria, Gemmatimonadetes, Firmicutes, Chlamydiae,Actinobacteria, Bacteroidetes*, OP11, JS1, OP3, OD1, BD1-5, TM6, SAR406. The relative abundace distribution is 150 *Cyanobacteria* (40.4%), 97 *Alphaproteobacteria* (26%), 39 *Gammaproteobacteria* (10.5%), 20 OD1 (5.4%), 10 SAR406 (3.7%). All groups less than 2.5% abundance were clustered together and are colored in purple. Circles represent known phyla and triangles represent candidate phyla.

Table S4.4. Horizontally transferred genes from archaeal and eukaryotes best matches for SAR406-CHDE genomes – complete

Genes with best BLAST matches to archaea or eukarya are displayed in this table including the product name annotation by IMG, the top hit species and the top hit BLAST match assessed by the DarkHorse

query_annotation	species Arch	besthit_annotation
SAR406-DSC14		
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
Thiamine pyrophosphate-requiring enzymes [acetolacta	Halococcus hamelinensis	acetolactate synthase
Aerobic-type carbon monoxide dehydrogenase, middle	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
Dienelactone hydrolase and related enzymes	Candidatus Nitrosopumilus sp. AR2	carboxymethylenebutenolidase
FOG: Ankyrin repeat	Metallosphaera yellowstonensis MK1	Chain A, Crystal Structure Of Engineered Protein. North
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
cysteine synthase (EC 2.5.1.47)	Candidatus Caldiarchaeum subterraneum	cysteine synthase A
Acyl dehydratase	Haloferax larsenii	dehydratase
Glycosyltransferases involved in cell wall biogenesis	Methanothermus fervidus DSM 2088	family 2 glycosyl transferase
Ferritin-like protein	Archaeoglobus sulfaticallidus PM70-1	Ferritin-like protein
Glutamate synthase domain 2	Candidatus Halobonum tyrrellensis	glutamate synthase
Predicted pyridoxal phosphate-dependent enzyme appa	Archaeoglobus veneficus SNP6	glutamine--scyllo-inositol transaminase
Predicted glycosylase	Ferroglobus placidus DSM 10642	glycosidase PH1107-related protein
Predicted glycosyltransferases	Methanocella conradii HZ254	glycosyltransferase
Predicted hydrolase (HAD superfamily)	Methanoregula boonei 6A8	HAD family hydrolase
Predicted phosphatases	Methanoculleus bourgensis MS2	hydrolase
ribonucleoside-diphosphate reductase class II (EC 1.17	Thermoplasmatales archaeon I-plasma	hypothetical protein
hypothetical protein	Thermoplasmatales archaeon SCGC AB-539-C06	hypothetical protein
Uncharacterized conserved protein	Archaeoglobus fulgidus DSM 4304	hypothetical protein
hypothetical protein	Methanosarcina acetivorans C2A	hypothetical protein
FOG: WD40-like repeat	Methanosaeta concilii GP6	hypothetical protein
Predicted membrane protein	Methanopyrus kandleri AV19	hypothetical protein
hypothetical protein	Methanocella conradii HZ254	hypothetical protein
Predicted integral membrane protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
Amidases related to nicotinamidase	Halococcus hamelinensis	isochorismatase
Amidases related to nicotinamidase	Haloquadratum walsbyi DSM 16790	isochorismatase
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Predicted kinase	Halogeometricum borinquense DSM 11551	kinase
DNA modification methylase	Methanoculleus bourgensis MS2	modification methylase
molybdopterin molybdochelatase (EC 2.10.1.1)	Aciduliprofundum sp. MAR08-339	molybdenum cofactor synthesis domain protein
N-acetyl sugar amidotransferase	Methanoregula formicica SMSP	N-acetyl sugar amidotransferase
N-acetylneuraminate synthase (EC 2.5.1.56)	Candidatus Nitrosoarchaeum limnia	N-acetylneuraminate synthase
3-methyladenine DNA glycosylase/8-oxoguanine DNA g	Methanoculleus bourgensis MS2	N-glycosylase/DNA lyase
NHL repeat	Methanosphaerula palustris E1-9c	NHL repeat containing protein
hypothetical protein	Thermoproteus uzoniensis 768-20	NUDIX hydrolase
HEAT repeats	Methanobacterium sp. SWAN-1	PBS lyase HEAT domain-containing protein
Phosphoenolpyruvate carboxykinase (GTP)	Pyrococcus yayanosii CH1	phosphoenolpyruvate carboxykinase
Metal-dependent hydrolases of the beta-lactamase sup	Thermococcus sp. 4557	phosphonate matabolism protein PhnP-like protein
Por secretion system C-terminal sorting domain	Pyrococcus yayanosii CH1	putative cysteinyl-tRNA synthetase
PQQ-like domain	Salinarchaeum sp. Harcht-Bsk1	pyrrolo-quinoline quinone
Saccharopine dehydrogenase and related proteins	Thermoplasmatales archaeon SCGC AB-539-N05	saccharopine dehydrogenase-like oxidoreductase
methyltransferase, FkbM family	Halalkalicoccus jeotgali B3	SAM-dependent methyltransferase
Subtilisin-like serine proteases	Methanolobus psychrophilus R15	subtilisin
endoribonuclease L-PSP	Pyrococcus sp. ST04	translation initiation inhibitor
UDP-N-acetylglucosamine 2-epimerase	Methanospirillum hungatei JF-1	UDP-N-acetylglucosamine 2-epimerase
SAR406-DSC15		
query_annotation	species Arch	besthit_annotation
Aerobic-type carbon monoxide dehydrogenase, middle	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
nucleoside diphosphate kinase (EC 2.7.4.6)	Thermoproteus uzoniensis 768-20	nucleoside-diphosphate kinase
molybdopterin molybdochelatase (EC 2.10.1.1)	Aciduliprofundum sp. MAR08-339	molybdenum cofactor synthesis domain protein
homoaconitate hydratase family protein/3-isopropylma	Aciduliprofundum sp. MAR08-339	homoaconitate hydratase family protein/3-isopropylmal
Uncharacterized conserved protein	Archaeoglobus fulgidus DSM 4304	hypothetical protein
Predicted pyridoxal phosphate-dependent enzyme appa	Archaeoglobus veneficus SNP6	glutamine--scyllo-inositol transaminase
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Protein of unknown function (DUF3179)	halophilic archaeon J07HX5	protein of unknown function (DUF3179)
conserved hypothetical protein	Haloarcula hispanica ATCC 33960	hypothetical protein HAH_1205
Glycosyltransferase	Haloarcula vallismortis	glycosyltransferase, type 1
Arginase/agmatinase/formimionoglutamate hydrolase,	Halobacterium sp. DL1	agmatinase
NAD dependent epimerase/dehydratase family	Halobacterium sp. NRC-1	GDP-D-mannose dehydratase
PA domain	Halococcus hamelinensis	peptidase M28
hypothetical protein	Halonotius sp. J07HN4	hypothetical protein
3-isopropylmalate dehydratase large subunit	Halonotius sp. J07HN4	aconitase A
Amidases related to nicotinamidase	Haloquadratum walsbyi DSM 16790	isochorismatase
PQQ-like domain	Salinarchaeum sp. Harcht-Bsk1	pyrrolo-quinoline quinone
HEAT repeats	Methanobacterium sp. SWAN-1	PBS lyase HEAT domain-containing protein
Predicted glycosyltransferases	Methanocella conradii HZ254	glycosyltransferase
Uncharacterized conserved protein	Methanocella conradii HZ254	hypothetical protein
DNA modification methylase	Methanoregula formicica SMSP	DNA modification methylase
putative efflux protein, MATE family	Methanosaeta concilii GP6	MATE efflux family protein
Predicted transcriptional regulator containing an HTH d	Methanosaeta harundinacea 6Ac	Putative transcriptional regulator
TIGR00725 family protein	Methanohalobium evestigatum Z-7303	hypothetical protein
hypothetical protein	Methanosarcina acetivorans C2A	hypothetical protein
Phosphoenolpyruvate carboxykinase (GTP)	Pyrococcus sp. ST04	phosphoenolpyruvate carboxykinase
3-isopropylmalate dehydrogenase (EC 1.1.1.85)	Pyrococcus sp. ST04	3-isopropylmalate dehydrogenase
endoribonuclease L-PSP	Pyrococcus sp. ST04	translation initiation inhibitor
Phosphoenolpyruvate carboxykinase (GTP)	Thermoplasmatales archaeon I-plasma	hypothetical protein
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
UDP-glucose 4-epimerase	Thermoplasmatales archaeon SCGC AB-539-N05	nucleoside-diphosphate-sugar epimerase
Saccharopine dehydrogenase and related proteins	Thermoplasmatales archaeon SCGC AB-539-N05	saccharopine dehydrogenase-like oxidoreductase
Predicted metal-sulfur cluster biosynthetic enzyme	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
hypothetical protein	uncultured marine crenarchaeote HF4000_APKG7F11	putative CoA-binding domain protein
Tryptophan synthase beta chain	Candidatus Nitrosoarchaeum limnia	tryptophan synthase subunit beta
N-acetylneuraminate synthase (EC 2.5.1.56)	Candidatus Nitrosoarchaeum limnia	N-acetylneuraminate synthase
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
Predicted integral membrane protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
NTP pyrophosphohydrolases including oxidative damag	Candidatus Nitrososphaera gargensis Ga9.2	NUDIX hydrolase

Table S4.4 Horizontally transferred genes from archaeal and eukaryotes best matches for SAR406-CHDE genomes -complete continued

SAR406-DSC16		
query_annotation	species Arch	besthit_annotation
hypothetical protein	Sulfolobus tokodaii str. 7	hypothetical protein
Xanthine and CO dehydrogenases maturation factor, X	Pyrobaculum sp. 1860	xanthine dehydrogenase accessory factor
Protein of unknown function (DUF3179)	halophilic archaeon J07HB67	protein of unknown function (DUF3179)
Predicted metal-dependent membrane protease	Haloferax sp. BAB2207	CAAX amino terminal protease family protein
hypothetical protein	Natrinema altunense	alkyl hydroperoxide reductase
Growth inhibitor	Natrinema pellirubrum DSM 15624	growth inhibitor
branched-chain amino acid aminotransferase, group I	Salinarchaeum sp. Harcht-Bsk1	branched-chain amino acid aminotransferase
NAD dependent epimerase/dehydratase family	Methanobacterium sp. Maddingley MBC34	nucleoside-diphosphate-sugar epimerase
Predicted ATPase involved in replication control, Cdc46,	Methanocaldococcus vulcanius M7	MCM family protein
5,10-methylene-tetrahydrofolate dehydrogenase/Methe	Methanoculleus sp. CAG:1088	bifunctional protein FolD
Response regulator containing CheY-like receiver, AAA-	Methanoregula formicica SMSP	response regulator with CheY-like receiver, AAA-type A
Response regulator receiver domain/Y_Y_Y domain/His	Methanosphaerula palustris E1-9c	PAS/PAC sensor hybrid histidine kinase [Methanosphae
Response regulator containing a CheY-like receiver don	Methanohalobium evestigatum Z-7303	response regulator receiver protein
hypothetical protein	Methanolobus tindarius	hypothetical protein
FOG: PKD repeat	Methanosarcina mazei Tuc01	cell surface protein
haloacid dehalogenase superfamily, subfamily IA, varia	Thermococcus barophilus MP	2-haloalkanoic acid dehalogenase
PEGA domain	Thermococcus sp. 4557	Serine/threonine protein kinase
Predicted integral membrane protein	Thermococcus sp. 4557	Serine/threonine protein kinase
ribonucleoside-diphosphate reductase class II (EC 1.17	Thermoplasmatales archaeon I-plasma	hypothetical protein
Por secretion system C-terminal sorting domain	uncultured marine crenarchaeote HF4000_APKG2O16	hypothetical protein
hypothetical protein	uncultured marine crenarchaeote HF4000_APKG2O16	hypothetical protein
SAR406-DSC17		
query_annotation	species Arch	besthit_annotation
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
ABC-type multidrug transport system, ATPase compone	Methanosarcina mazei Tuc01	ABC transporter ATP-binding protein
ABC-type dipeptide/oligopeptide/nickel transport syste	Halovivax ruber XH-70	ABC-type dipeptide/oligopeptide/nickel transport syster
Thiamine pyrophosphate-requiring enzymes [acetolacta	Halococcus hamelinensis	acetolactate synthase
His Kinase A (phospho-acceptor) domain/PAS fold/Histi	Methanococcus maripaludis X1	ATPase-like ATP-binding protein
Plastocyanin	Candidatus Nitrososphaera gargensis Ga9.2	blue (Type 1) copper domain-containing protein
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
Uncharacterized membrane protein	Candidatus Caldiarchaeum subterraneum	conserved hypothetical protein
Cysteine sulfinate desulfinase/cysteine desulfurase and	Methanobacterium sp. SWAN-1	Cysteine desulfurase
cysteine synthase (EC 2.5.1.47)	Candidatus Caldiarchaeum subterraneum	cysteine synthase A
Phosphoglycerate dehydrogenase and related dehydrog	Methanoregula formicica SMSP	D-3-phosphoglycerate dehydrogenase
FAD/FMN-containing dehydrogenases	Candidatus Nitrososphaera gargensis Ga9.2	FAD linked oxidase-like protein
Aerobic-type carbon monoxide dehydrogenase, middle	Desulfurococcus kamchatkensis 1221n	FAD-binding molybdopterin dehydrogenase
Predicted glycosylase	Ferroglobus placidus DSM 10642	glycosidase PH1107-related protein
Glycosyltransferase	Methanobacterium sp. Maddingley MBC34	glycosyltransferase
Glycosyl transferase 4-like	Methanothermobacter thermautotrophicus CaT2	glycosyltransferase
haloacid dehalogenase superfamily, subfamily IA, varia	Methanothermococcus okinawensis IH1	HAD superfamily hydrolase
FOG: HEAT repeat	Methanococcoides burtonii DSM 6242	HEAT repeat-containing PBS lyase
homoaconitate hydratase family protein/3-isopropylma	Aciduliprofundum sp. MAR08-339	homoaconitate hydratase family protein/3-isopropylmal
Predicted phosphatases	Methanoculleus bourgensis MS2	hydrolase
Uncharacterized conserved protein	Archaeoglobus fulgidus DSM 4304	hypothetical protein
hypothetical protein	uncultured marine crenarchaeote HF4000_APKG2O16	hypothetical protein
hypothetical protein	Archaeoglobus sulfaticallidus PM70-1	hypothetical protein
Protein of unknown function (DUF3179)	Halogeometricum borinquense DSM 11551	hypothetical protein
hypothetical protein	Halorhabdus utahensis DSM 12940	hypothetical protein
Protein of unknown function (DUF1670)	Methanosarcina acetivorans C2A	hypothetical protein
TPR repeat	Methanosaeta harundinacea 6Ac	hypothetical protein
Predicted membrane protein	Methanopyrus kandleri AV19	hypothetical protein
Predicted DNA alkylation repair enzyme	Methanosarcina mazei Go1	hypothetical protein
TIGR00725 family protein	Methanohalophilus mahii DSM 5219	hypothetical protein
AAA ATPase domain	Methanococcus maripaludis C7	hypothetical protein
Transcriptional regulators of sugar metabolism	Methanosarcina mazei Tuc01	hypothetical protein
Amidases related to nicotinamidase	Haloquadratum walsbyi DSM 16790	isochorismatase
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
hypothetical protein	Halorhabdus tiamatea SARL4B	major facilitator superfamily MFS_1
Malate/lactate dehydrogenases	Methanohalobium evestigatum Z-7303	malate dehydrogenase
putative efflux protein, MATE family	Methanosaeta concilii GP6	MATE efflux family protein
DNA modification methylase	Methanoculleus bourgensis MS2	modification methylase
molybdopterin molybdochelatase (EC 2.10.1.1)	Aciduliprofundum sp. MAR08-339	molybdenum cofactor synthesis domain protein
Serine acetyltransferase	Archaeoglobus veneficus SNP6	N-acetylglucosamine-1-phosphateuridyltransferase
NHL repeat	Methanosphaerula palustris E1-9c	NHL repeat containing protein
NTP pyrophosphohydrolases including oxidative damag	Candidatus Nitrososphaera gargensis Ga9.2	NUDIX hydrolase
Raf kinase inhibitor-like protein, YbhB/YbcL family	Methanosphaerula palustris E1-9c	PEBP family protein
hypothetical protein	Methanosarcina mazei Go1	phosphate ABC transporter permease
Phosphoenolpyruvate carboxykinase (GTP)	Thermococcus sp. 4557	phosphoenolpyruvate carboxykinase
Por secretion system C-terminal sorting domain	uncultured marine crenarchaeote HF4000_APKG2O16	putative fibronectin type III domain protein
Predicted restriction endonuclease	Methanolobus tindarius	putative restriction endonuclease
Response regulator containing CheY-like receiver, AAA-	Methanoregula formicica SMSP	response regulator with CheY-like receiver, AAA-type A
Stress responsive A/B Barrel Domain	Methanococcus maripaludis X1	stress responsive alpha-beta barrel domain-containing
hypothetical protein	Thermoplasmatales archaeon Gpl	sulfide-quinone reductase related protein
endoribonuclease L-PSP	Pyrococcus sp. ST04	translation initiation inhibitor
NAD dependent epimerase/dehydratase family	Pyrococcus sp. ST04	UDP-glucose 4-epimerase
hypothetical protein	Candidatus Nitrosoarchaeum koreensis	Zn-ribbon protein

Table S4.4 Horizontally transferred genes from archaeal and eukaryotes best matches for SAR406-CHDE genomes -complete continued

SAR406-DSC18		
query_annotation	species Arch	besthit_annotation
Aerobic-type carbon monoxide dehydrogenase, middle	Sulfolobus tokodaii str. 7	carbon-monoxide dehydrogenase middle subunit
homoaconitate hydratase family protein/3-isopropylma	Aciduliprofundum sp. MAR08-339	homoaconitate hydratase family protein/3-isopropylmal
Asparagine synthase (glutamine-hydrolyzing)	Aciduliprofundum sp. MAR08-339	asparagine synthase, glutamine-hydrolyzing
Asparagine synthase (glutamine-hydrolyzing)	Aciduliprofundum sp. MAR08-339	asparagine synthase, glutamine-hydrolyzing
Uncharacterized conserved protein	Archaeoglobus fulgidus DSM 4304	hypothetical protein
Acetyl-CoA acetyltransferase	Archaeoglobus veneficus SNP6	acetyl-CoA acetyltransferase
ABC-type nitrate/sulfonate/bicarbonate transport syste	Archaeoglobus veneficus SNP6	ABC transporter permease
isocitrate dehydrogenase (NADP) (EC 1.1.1.42)	Archaeoglobus veneficus SNP6	isocitrate dehydrogenase
Predicted glycosylase	Ferroglobus placidus DSM 10642	glycosidase PH1107-related protein
Uncharacterized conserved protein	haloarchaeon 3A1_DGR	hypothetical protein
Uncharacterized conserved protein	Halalkalicoccus jeotgali B3	phenylacetic acid catabolic family protein
Phosphotransacetylase	Halarchaeum acidiphilum	hypothetical protein
Acyl dehydratase	Haloferax larsenii	dehydratase
Protein of unknown function (DUF4242)	Haloferax mediterranei	putative gualylate cyclase protein
FOG: WD40-like repeat	Halonotius sp. J07HN4	WD40 repeat protein
hypothetical protein	Halorhabdus tiamatea SARL4B	major facilitator superfamily MFS_1
Uncharacterized conserved protein	Natrialba hulunbeirensis	phenylacetate-CoA oxygenase subunit PaaI
Membrane protein involved in the export of O-antigen a	Methanobacterium sp. AL-21	polysaccharide biosynthesis protein
Cysteine sulfinate desulfinase/cysteine desulfurase and	Methanobacterium sp. SWAN-1	Cysteine desulfurase
HEAT repeats	Methanobacterium sp. SWAN-1	PBS lyase HEAT domain-containing protein
haloacid dehalogenase superfamily, subfamily IA, varia	Methanocaldococcus infernus ME	HAD superfamily (subfamily IA) hydrolase, TIGR02253
hypothetical protein	Methanotorris igneus Kol 5	hypothetical protein
Predicted glycosyltransferases	Methanocella conradii HZ254	glycosyltransferase
3-methyladenine DNA glycosylase/8-oxoguanine DNA g	Methanoculleus bourgensis MS2	N-glycosylase/DNA lyase
Predicted phosphatases	Methanoculleus bourgensis MS2	hydrolase
Predicted DNA alkylation repair enzyme	Methanoregula boonei 6A8	hypothetical protein
Response regulator containing CheY-like receiver, AAA-	Methanoregula formicica SMSP	response regulator with CheY-like receiver, AAA-type A
ATP-dependent exoDNAse (exonuclease V) beta subuni	Methanoregula formicica SMSP	ATP-dependent exonuclase V beta subunit, helicase and
Gluconolactonase	Methanosphaerula palustris E1-9c	PKD domain-containing protein [Methanosphaerula palu
Por secretion system C-terminal sorting domain	Methanosaeta concilii GP6	hypothetical protein
FOG: HEAT repeat	Methanococcoides burtonii DSM 6242	HEAT repeat-containing PBS lyase
Threonine dehydrogenase and related Zn-dependent de	Methanolobus psychrophilus R15	oxidoreductase
Predicted metal-sulfur cluster biosynthetic enzyme	Methanosarcina acetivorans C2A	hypothetical protein
hypothetical protein	Methanosarcina mazei Go1	phosphate ABC transporter permease
Uncharacterized membrane protein, putative virulence	Methanosarcina mazei Tuc01	polysaccharide biosynthesis protein
Predicted membrane protein	Methanopyrus kandleri AV19	hypothetical protein
hypothetical protein	Pyrococcus horikoshii OT3	hypothetical protein
hypothetical protein	Pyrococcus horikoshii OT3	hypothetical protein
Uncharacterized proteins of the AP superfamily	Thermococcus barophilus MP	hypothetical protein
putative oligopeptide transporter, OPT family	Thermococcus onnurineus NA1	oligopeptide transporter
ribonucleoside-diphosphate reductase class II (EC 1.17	Thermoplasmatales archaeon I-plasma	hypothetical protein
hypothetical protein	Thermoplasmatales archaeon SCGC AB-539-C06	hypothetical protein
Isocitrate/isopropylmalate dehydrogenase	Thermoplasmatales archaeon SCGC AB-539-C06	isocitrate/isopropylmalate dehydrogenase
Saccharopine dehydrogenase and related proteins	Thermoplasmatales archaeon SCGC AB-539-N05	saccharopine dehydrogenase-like oxidoreductase
NTP pyrophosphohydrolases including oxidative damag	Candidatus Caldiarchaeum subterraneum	NUDIX hydrolase
Predicted metal-sulfur cluster biosynthetic enzyme	Cenarchaeum symbiosum A	metal-sulfur cluster biosynthetic enzyme
Predicted integral membrane protein	Candidatus Nitrosopumilus sp. AR2	hypothetical protein
SAR406-DSC19		
query_annotation	species Arch	besthit_annotation
pterin-4-alpha-carbinolamine dehydratase (EC 4.2.1.96	Candidatus Caldiarchaeum subterraneum	4a-hydroxytetrahydrobiopterin dehydratase
cob(I)yrinic acid a,c-diamide adenosyltransferase (EC 2	Candidatus Nitrosopumilus salaria	cob(I)yrinic acid a,c-diamide adenosyltransferase
Predicted glycosylase	Ferroglobus placidus DSM 10642	glycosidase PH1107-related protein
Predicted glycosyltransferases	Methanobrevibacter ruminantium M1	glycosyl transferase GT2 family
hypothetical protein	Methanobacterium sp. Maddingley MBC34	hypothetical protein
hypothetical protein	Methanolinea tarda	hypothetical protein
hypothetical protein	Methanolinea tarda	hypothetical protein
Phosphoenolpyruvate carboxykinase (GTP)	Thermoplasmatales archaeon I-plasma	hypothetical protein
Predicted pyridoxal phosphate-dependent enzyme appa	Ferroplasma acidarmanus fer1	hypothetical protein
TIGR00725 family protein	Methanohalophilus mahii DSM 5219	hypothetical protein
Protein of unknown function (DUF3179)	Nitrosopumilus maritimus SCM1	hypothetical protein
Amidases related to nicotinamidase	Haloquadratum walsbyi DSM 16790	isochorismatase
Predicted metal-sulfur cluster biosynthetic enzyme	Cenarchaeum symbiosum A	metal-sulfur cluster biosynthetic enzyme
molybdopterin molybdochelatase (EC 2.10.1.1)	Aciduliprofundum sp. MAR08-339	molybdenum cofactor synthesis domain protein
hypothetical protein	Haloferax sp. BAB2207	NhaC-type sodium/hydrogen antiporter, partial
nucleotide sugar dehydrogenase	Methanoregula formicica SMSP	nucleotide sugar dehydrogenase
ADP-ribose pyrophosphatase	Methanobacterium sp. AL-21	NUDiX hydrolase
hypothetical protein	Aciduliprofundum sp. MAR08-339	NurA domain-containing protein
Uncharacterized conserved protein	Halalkalicoccus jeotgali B3	phenylacetic acid catabolic family protein
Uncharacterized conserved protein	Salinarchaeum sp. Harcht-Bsk1	phenylacetic acid degradation protein PaaC
Phosphoenolpyruvate carboxykinase (GTP)	Pyrococcus sp. ST04	phosphoenolpyruvate carboxykinase
Plasmid pRiA4b ORF-3-like protein	Thermoplasmatales archaeon A-plasma	Plasmid pRiA4b ORF-3 family protein
ATP-dependent 26S proteasome regulatory subunit	Methanoregula boonei 6A8	proteasome-activating nucleotidase
Predicted ATPase	Aciduliprofundum sp. MAR08-339	putative ATPase
Protein of unknown function (DUF4242)	Haloferax mediterranei	putative gualylate cyclase protein
HI0933-like protein	Methanosarcina barkeri	Uncharacterized protein in nifH2 5region
FOG: CheY-like receiver	Methanolobus psychrophilus R15	response regulator receiver protein
Uncharacterized protein conserved in bacteria	Methanosarcina mazei Go1	serine/threonine protein kinase
Site-specific recombinase XerD	Aciduliprofundum sp. MAR08-339	site-specific recombinase XerD
Stress responsive A/B Barrel Domain	Methanococcus maripaludis X1	stress responsive alpha-beta barrel domain-containing
Threonine synthase	Methanocaldococcus vulcanius M7	threonine synthase
Threonine synthase	Candidatus Caldiarchaeum subterraneum	threonine synthase, partial
Threonine synthase	Natronococcus amylolyticus	threonine synthase, partial
FOG: WD40-like repeat	halophilic archaeon J07HX64	WD40 repeat protein

Table S4.4 Horizontally transferred genes from archaeal and eukaryotes best matches for SAR406-CHDE genomes -complete continued

SAR406-DSC14		
Ribosomal protein L5	Chlorella sp. ArM0029B	50S ribosomal protein L5 (chloroplast)
Short-chain dehydrogenases of various substrate speci	Aspergillus flavus NRRL3357	estradiol 17 beta-dehydrogenase
Dolichol kinase	Dictyostelium discoideum AX4	hypothetical protein
Iron-binding zinc finger CDGSH type	Polysphondylium pallidum PN500	hypothetical protein
FG-GAP repeat	Thalassiosira oceanica	hypothetical protein
NADH:ubiquinone oxidoreductase subunit 5 (chain L)/M	Fagraea caudata	NADH dehydrogenase subunit F, partial (chloroplast)
Ornithine carbamoyltransferase	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
Repeat domain in Vibrio, Colwellia, Bradyrhizobium and	Micromonas sp. RCC299	predicted protein
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein
SAR406-DSC15		
query_annotation	species Euk	besthit_annotation
Ribosomal protein L5	Chlorella sp. ArM0029B	50S ribosomal protein L5 (chloroplast)
hypothetical protein	Taenia asiatica	ATPase
Branched-chain amino acid aminotransferase/4-amino-	Penicillium digitatum PHI26	D-alanine aminotransferase
hypothetical protein	Saccharomyces cerevisiae AWRI796	hypothetical protein
Glutamine amidotransferases class-II	Moniliophthora perniciosa FA553	hypothetical protein
Iron-binding zinc finger CDGSH type	Polysphondylium pallidum PN500	hypothetical protein
Pyruvate/2-oxoglutarate dehydrogenase complex, dihy	Tremella fuciformis	pyridine redox protein
Cell division protein FtsI/penicillin-binding protein 2	Trypanosoma congolense IL3000	unnamed protein product
SAR406-DSC16		
query_annotation	species Euk	besthit_annotation
ABC-type transport system involved in Fe-S cluster ass	Phillyrea latifolia	ABC transporter precursor
Uncharacterized protein involved in tolerance to divalen	Toxoplasma gondii ME49	divalent cation tolerance protein, CutA1 family protein
Glucose-6-phosphate 1-dehydrogenase	Urodus sp. CR16	glucose phosphate dehydrogenase
Predicted oxidoreductases (related to aryl-alcohol dehy	Eutrema salsugineum	hypothetical protein
3-dehydroquinate synthase (EC 4.2.3.4)	Nematostella vectensis	hypothetical protein
Dolichol kinase	Polysphondylium pallidum PN500	hypothetical protein
Cytosine/uracil/thiamine/allantoin permeases	Prunus persica	hypothetical protein
Ornithine carbamoyltransferase	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
Iodothyronine deiodinase	Nematostella vectensis	predicted protein
Prolyl oligopeptidase family	Nematostella vectensis	predicted protein
Glutamate dehydrogenase/leucine dehydrogenase	Fragaria vesca subsp. vesca	PREDICTED: glutamate dehydrogenase 2-like
S-adenosylhomocysteine hydrolase	Polyporales sp. KUC9061	putative S-adenosyl-L-homocysteine hydrolase
SAR406-DSC17		
query_annotation	species Euk	besthit_annotation
Ribosomal protein L5	Chlorella sp. ArM0029B	50S ribosomal protein L5 (chloroplast)
Copper chaperone	Antilocapra americana	ATPase
Response regulator containing CheY-like receiver, AAA-	Ricinus communis	conserved hypothetical protein
Male sterility protein	Fusarium oxysporum f. sp. cubense race 1	Fatty acyl-CoA reductase 1
Por secretion system C-terminal sorting domain	Aureococcus anophagefferens	hypothetical protein
Putative silver efflux pump	Capitella teleta	hypothetical protein
Dolichol kinase	Dictyostelium discoideum AX4	hypothetical protein
Exonuclease III	Genlisea aurea	hypothetical protein
ornithine carbamoyltransferase (EC 2.1.3.3)	Ostreococcus lucimarinus CCE9901	Ornithine carbamoyltransferase
FG-GAP repeat	Micromonas pusilla CCMP1545	predicted protein
FG-GAP repeat	Micromonas sp. RCC299	predicted protein
FG-GAP repeat	Micromonas sp. RCC299	predicted protein
hypothetical protein	Amphimedon queenslandica	adenylyltransferase and sulfurtransferase MOCS3-like
Subtilisin-like serine proteases	Lagenidium giganteum	subtilisin-like serine protease
SAR406-DSC18		
query_annotation	species Euk	besthit_annotation
Ribosomal protein L5	Chlorella sp. ArM0029B	50S ribosomal protein L5 (chloroplast)
5-formyltetrahydrofolate cyclo-ligase	Exophiala dermatitidis NIH/UT8656	5-formyltetrahydrofolate cyclo-ligase
hypothetical protein	Ricinus communis	conserved hypothetical protein
Heme/copper-type cytochrome/quinol oxidases, subuni	Microplitis sp. jft91	cytochrome oxidase subunit 1
hypothetical protein	Batrachochytrium dendrobatidis JAM81	hypothetical protein
Dolichol kinase	Dictyostelium discoideum AX4	hypothetical protein
Cysteine synthase	Meyerozyma guilliermondii ATCC 6260	hypothetical protein
Iron-binding zinc finger CDGSH type	Polysphondylium pallidum PN500	hypothetical protein
Flp pilus assembly protein TadD, contains TPR repeats	Trichoplax adhaerens	hypothetical protein
UDP-glucose 4-epimerase	Perkinsus marinus ATCC 50983	nad dependent epimerase/dehydratase
ornithine carbamoyltransferase (EC 2.1.3.3)	Coccomyxa subellipsoidea C-169	ornithine carbamoyltransferase
hypothetical protein	Physcomitrella patens	predicted protein
Acetyl-CoA carboxylase, carboxyltransferase componer	Gallus gallus	probable methylcrotonoyl-CoA carboxylase beta chain
SAR406-DSC19		
query_annotation	species Euk	besthit_annotation
Ribosomal protein L5	Chlorella sp. ArM0029B	50S ribosomal protein L5 (chloroplast)
Copper chaperone	Martes americana	ATP7A
Glutamate dehydrogenase/leucine dehydrogenase	Hypophthalmichthys nobilis	glutamate dehydrogenase 1
hypothetical protein	Monosiga brevicollis MX1	hypothetical protein
Receptor L domain	Monosiga brevicollis MX1	hypothetical protein
hypothetical protein	Monosiga brevicollis MX1	hypothetical protein
Por secretion system C-terminal sorting domain	Aureococcus anophagefferens	hypothetical protein
Dolichol kinase	Dictyostelium discoideum AX4	hypothetical protein
FG-GAP repeat	Thalassiosira oceanica	hypothetical protein
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein
Por secretion system C-terminal sorting domain	Micromonas sp. RCC299	predicted protein
Tetratricopeptide repeat	Nematostella vectensis	predicted protein
Phosphoenolpyruvate synthase/pyruvate phosphate dik	Schistosoma mansoni	pyruvate phosphate dikinase chloroplast
hypothetical protein	Macrophomina phaseolina MS6	Vitamin B6 biosynthesis protein

REFRENCES

Allers E, Wright JJ, Konwar KM, Howes CG, Beneze E, Hallam SJ, Sullivan MB (2013). Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean. *ISME J* **7:** 256-268.

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

Anand A, Satyanarayana T (2012). Applicability of carboxydotrophic bacterial carbon monoxide dehydrogenase (CODH) in carbon sequestration and bioenergy generation. *Journal of Scientific & Industrial Research* **71:** 381-384.

Arístegui J, Duarte CM, Gasol JM, Herndl GJ (2009-04-16). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography* **54**.

Atack JM, Srikhanta YN, Djoko KY, Welch JP, Hasri NHM, Steichen CT, Vanden H, Rachel N, Grimmond, SM, Othman D, Kappler U, Apicella MA, Jennings, MP, Edwards JL, McEwan AG (2013). Characterization of an ntrX Mutant of Neisseria gonorrhoeae Reveals a Response Regulator That Controls Expression of Respiratory Enzymes in Oxidase-Positive Proteobacteria. *Journal of Bacteriology* **195:** 2632-2641.

Atichartpongkul S, Loprasert S, Vattanaviboon P, Whangsuk W, Helmann JD, Mongkolsuk S (2001). Bacterial Ohr and OsmC paralogues define two protein families with distinct functions and patterns of expression. *Microbiology* **147:** 1775-1782.

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov A Lesin, VM, Nikolenko, SI, Pham, S, Prjibelski, AD, Pyshkin, AV, Sirotkin, AV, Vyahhi, N, Tesler, G, Alekseyev, MA, Pevzner, PA (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19:** 455-477.

Barrangou R, Horvath P (2012). CRISPR: New Horizons in Phage Resistance and Strain Identification. *Annual Review of Food Science and Technology* **3:** 143-162.

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S (2012). The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol* **13:** R39.

Bowman JP, McCuaig RD (2003). Biodiversity, Community Structural Shifts, and Biogeography of Prokaryotes within Antarctic Continental Shelf Sediment. *Applied and Environmental Microbiology* **69:** 2463-2483.

Brown MV, Philip GK, Bunge JA, Smith MC, Bissett A, Lauro FM, Fuhrman JA,

Donachie SP (2009). Microbial community structure in the North Pacific ocean. *The ISME journal* **3:** 1374-1386.

Calamita G (2000). The Escherichia coli aquaporin-Z water channel. *Molecular Microbiology* **37:** 254-262.

Coates JD, Ellis DJ, Gaw CV, Lovley DR (1999). Geothrix fermentans gen. nov., sp. nov., a novel Fe(III)-reducing bacterium from a hydrocarbon-contaminated aquifer. *International Journal of Systematic and Evolutionary Microbiology* **49:** 1615-1622.

Crump BC, Peranteau C, Beckingham B, Cornwell JC (2007). Respiratory Succession and Community Succession of Bacterioplankton in Seasonally Anoxic Estuarine Waters. *Applied and Environmental Microbiology* **73:** 6802-6810.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72:** 5069-5072.

DiPippo JL, Nesbø CL, Dahle H, Doolittle WF, Birkland N-K, Noll KM (2009). Kosmotoga olearia gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. *International Journal of Systematic and Evolutionary Microbiology* **59:** 2991-3000.

Durbin AM, Teske A (2010). Sediment-associated microdiversity within the Marine Group I Crenarchaeota. *Environmental Microbiology Reports* **2:** 693-703.

Durbin AM, Teske A (2011). Microbial diversity and stratification of South Pacific abyssal marine sediments. *Environmental Microbiology* **13:** 3219-3234.

Eloe E, Fadrosh D, Novotny M, Allen L, Kim M, Lombardo M Yee-Greenbaum, J, Yooseph, S, Allen, EE, Lasken, R, Williamson, SJ, Bartlett, DH (2011a). Going Deeper: Metagenome of a Hadopelagic Microbial Community. *Plos One* **6**.

Eloe E, Shulse C, Fadrosh D, Williamson S, Allen E, Bartlett D (2011b). Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environmental Microbiology Reports* **3:** 449-458.

Ferry JG (1995). Co Dehydrogenase. *Annual Review of Microbiology* **49:** 305-333.

Fuchs BM, Woebken D, Zubkov MV, Burkill P, Amann R (2005). Molecular identification of picoplankton populationsin contrasting waters of the Arabian Sea. *Aquatic Microbial Ecology* **39:** 145-157.

Fuhrman JA, McCallum K, Davis AA (1993). Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Applied and Environmental Microbiology* **59:** 1294-1302.

Fujioka K, Okino K, Kanamatsu T, Ohara Y (2002). Morphology and origin of the Challenger Deep in the Southern Mariana Trench. *Geophysical Research Letters* **29:** 10-11-10-14.

Gallagher JM, Carton MW, Eardly DF, Patching JW (2004). Spatio-temporal variability and diversity of water column prokaryotic communities in the eastern North Atlantic. *FEMS Microbiology Ecology* **47:** 249-262.

Gordon DA, Giovannoni SJ (1996). Detection of stratified microbial populations related to Chlorobium and Fibrobacter species in the Atlantic and Pacific oceans. *Applied and Environmental Microbiology* **62:** 1171-1177.

Griesbeck C, HAUSKA G, Schütz M (2000). *Biological sulfide oxidation: Sulfide-quinone reductase (SQR), the primary reaction*.

Gupta RS (2004). The Phylogeny and Signature Sequences Characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. *Critical Reviews in Microbiology* **30:** 123-143.

Hardy K (2013). Greg MacEachern, Edgetech, Inc., West Wareham, MA John Head, Prevco, Fountain Hills, AZ Larry Herbst, Earthlight Communications, Pasadena, CA Untethered free vehicles are absolutely the most cost-effective way to get to mid-water.

Hille R (2005). Molybdenum-containing hydroxylases. *Archives of Biochemistry and Biophysics* **433:** 107-116.

Horvath P, Barrangou R (2010). CRISPR/Cas, the Immune System of Bacteria and Archaea. *Science* **327:** 167-170.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA Sogin ML (2007). Microbial Population Structures in the Deep Marine Biosphere. *Science* **318:** 97-100.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biol* **8:** R143.

Huse SM, Welch DBM, Voorhis A, Shipunova A, Morrison HG, Eren AM, Sogin ML (2014). VAMPS: a website for visualization and analysis of microbial population structures. *BMC bioinformatics* **15:** 41.

Jones AC, Monroe EA, Podell S, Hess WR, Klages S, Esquenazi E, Niessen, S., Hoover,

H, Rothmann, M, Lasken, RS, Yates, JR, Reinhardt, R, Kube, M, Burkart, MD, Allen, EE, Dorrestein, PC, Gerwick, WH, Gerwick, Lena (2011). Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium Lyngbya majuscula. *Proceedings of the National Academy of Sciences* **108:** 8815-8820.

Kappes RM, Kempf B, Kneip S, Boch J, Gade J, Meier-Wagner J, Bremer E (1999). Two evolutionarily closely related ABC transporters mediate the uptake of choline for synthesis of the osmoprotectant glycine betaine in Bacillus subtilis. *Molecular Microbiology* **32:** 203-216.

Kato S, Kobayashi C, Kakegawa T, Yamagishi A (2009). Microbial communities in iron-silica-rich microbial mats at deep-sea hydrothermal fields of the Southern Mariana Trough. *Environmental Microbiology* **11:** 2094-2111.

Kempf B, Bremer E (1995). OpuA, an Osmotically Regulated Binding Protein-dependent Transport System for the Osmoprotectant Glycine Betaine in Bacillus subtilis. *Journal of Biological Chemistry* **270:** 16701-16713.

King GM, Weber CF (2007). Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nature Reviews Microbiology* **5:** 107-118.

Knowles R (1982). Denitrification. *Microbiological Reviews* **46:** 43-70.

Kredich NM, Tomkins GM (1966). The Enzymic Synthesis of l-Cysteine in Escherichia coli and Salmonella typhimurium. *Journal of Biological Chemistry* **241:** 4955-4965.

Lauro FM, Bartlett DH (2008). Prokaryotic lifestyles in deep sea habitats. *Extremophiles* **12:** 15-25.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O (2012). Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31:** 1-46.

Lesniak J, Barton WA, Nikolov DB (2003). Structural and functional features of the Escherichia coli hydroperoxide resistance protein OsmC. *Protein Science* **12:** 2838-2843.

Markowitz V, Chen I, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner, A Huang, JH, Woyke, T, Huntemann, M, Anderson, I, Billis, K, Varghese, N, Mavromatis, K, Pati, A, Ivanova, NN, Kyrpides, NC (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42:** D560-D567.

Martin D, Bartlett D, Roberts M (2002). Solute accumulation in the deep-sea bacterium

Photobacterium profundum. *Extremophiles* **6:** 507-514.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi, N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences* **110:** E2390-E2399.

Meyer O, Rajagopalan KV (1984). Molybdopterin in carbon monoxide oxidase from carboxydotrophic bacteria. *Journal of Bacteriology* **157:** 643-648.

Miroshnichenko ML, L'Haridon S, Jeanthon C, Antipov AN, Kostrikina NA, Tindall BJ, Schumann P, Spring S, Stackebrandt E, Bonch-Osmolovskaya EA (2003). Oceanithermus profundus gen. nov., sp. nov., a thermophilic, microaerophilic, facultatively chemolithoheterotrophic bacterium from a deep-sea hydrothermal vent. *International Journal of Systematic and Evolutionary Microbiology* **53:** 747-752.

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36:** 760-766.

Naponelli V, Noiriel A, Ziemak MJ, Beverley SM, Lye L-F, Plume AM, Botella JR, Loizeau K, Ravanel S, Rébeillé F, de Crécy-Lagard V, Hanson AD (2008). Phylogenomic and Functional Analysis of Pterin-4a-Carbinolamine Dehydratase Family (COG2154) Proteins in Plants and Microorganisms. *Plant Physiology* **146:** 1515-1527.

Nesbø C, Bradnan D, Adebusuyi A, Dlutek M, Petrus A, Foght J, Doolittle WF, Noll K (2012). Mesotoga prima gen. nov., sp. nov., the first described mesophilic species of the Thermotogales. *Extremophiles* **16:** 387-393.

Nunoura T, Takaki Y, Kazama H, Hirai M, Ashi J, Imachi H, Takai K (2012). Microbial Diversity in Deep-sea Methane Seep Sediments Presented by SSU rRNA Gene Tag Sequencing. *Microbes and Environments* **27:** 382-390.

Nübel T, Klughammer C, Huber R, Hauska G, Schütz M (2000). Sulfide:quinone oxidoreductase in membranes of the hyperthermophilic bacterium Aquifex aeolicus (VF5). *Archives of Microbiology* **173:** 233-244.

Ochman H, Lawrance JG, Groisman E (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405:** 299.

Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ (2011). Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiology and Molecular Biology Reviews* **75:** 361-422.

Pace NR (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science* **276:** 734-740.

Park S-C, Pham BP, Van Duyet L, Jia B, Lee S, Yu R, Woo Han S, Yang, JK, Hahm K-S, Cheong GW (2008). Structural and functional characterization of osmotically inducible protein C (OsmC) from Thermococcus kodakaraensis KOD1. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1784:** 783-788.

Pedrós-Alió C (2006). Marine microbial diversity: can it be determined? *Trends in Microbiology* **14:** 257-263.

Pedrós-Alió C (2011). The Rare Bacterial Biosphere. *Annual Review of Marine Science* **4:** 449-466.

Peschek G, Löffelhardt W, Schmetterer G, Shahak Y, Schütz M, Bronstein M, Griesbeck C, Hauska G, Padan E (1999). Sulfide-Dependent Anoxygenic Photosynthesis in Prokaryotes. *The Phototrophic Prokaryotes*. Springer US. pp 217-228.

Pikuta EV. (2011).Overview of Archaea. *SPIE Optical Engineering+ Applications*.

Podell S, Gaasterland T (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology* **8**.

Podosokorskaya OA, Kadnikov VV, Gavrilov SN, Mardanov AV, Merkel AY, Karnachuk OV, Ravin NV, Bonch-Osmolovskaya EA, Kublanov IV (2013). Characterization of Melioribacter roseus gen. nov., sp. nov., a novel facultatively anaerobic thermophilic cellulolytic bacterium from the class Ignavibacteria, and a proposal of a novel bacterial phylum Ignavibacteriae. *Environmental Microbiology* **15:** 1759-1771.

Price M, Dehal P, Arkin A (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26:** 1641-1650.

Pruesse E, Peplies J, Glockner F (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28:** 1823-1829.

Quince C, Curtis TP, Sloan WT (2008). The rational exploration of microbial diversity. *The ISME journal* **2:** 997-1006.

Rappé MS, Giovannoni SJ (2003). The uncultured microbial majority. *Annual Review of Microbiology* **57:** 369-394.

Raux E, Schubert HL, Warren MJ (2000). Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum. *Cellular and Molecular Life Sciences CMLS* **57:** 1880-1893.

Ravin NV, Mardanov AV, Beletsky AV, Kublanov IV, Kolganova TV, Lebedinsky AV, Chernyh NA, Bonch-Osmolovskaya EA, Skryabin KG (2009). Complete Genome Sequence of the Anaerobic, Protein-Degrading Hyperthermophilic Crenarchaeon Desulfurococcus kamchatkensis. *Journal of Bacteriology* **191:** 2371-2379.

Reinthaler T, van Aken HM, Herndl GJ (2010). Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. *Deep Sea Research Part II: Topical Studies in Oceanography* **57:** 1572-1580.

Rinke C, Schwientek P, Sczyrba A, Ivanova N, Anderson I, Cheng J, Darling, A, Malfatti, S, Swan, BK, Gies, EA, Dodsworth, JA, Hedlund, BP, Tsiamis, G, Sievert, SM, Liu, WT, Eisen, JA, Hallam, SJ, Kyrpides, NC, Stepanauskas, R, Rubin, EM, Hugenholtz, P, Woyke, T (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Rinke C, Lee J, Nath N, Goudeau D, Thompson B, Poulton N, Dmitrieff E, Malmstrom R, Stepanauskas R, Woyke T (2014). Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nature protocols* **9:** 1038-1048.

Robinson CR, Sligar SG, Michael L. Johnson GKA (1995). [18] Hydrostatic and osmotic pressure as tools to study macromolecular recognition. *Methods in Enzymology*. Academic Press. pp 395-427.

Saier JMH (2000). Families of transmembrane transporters selective for amino acids and their derivatives. *Microbiology* **146:** 1775-1795.

Schauer R, Bienhold C, Ramette A, Harder J (2009). Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME J* **4:** 159-170.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75:** 7537-7541.

Smedile F, Messina E, La Cono V, Tsoy O, Monticelli L, Borghini M, Giuliano, L, Golyshin, PN, Mushegian, A, Yakimov, MM (2013). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of Mediterranean Sea, Matapan-Vavilov Deep. *Environmental Microbiology* **15:** 167-182.

Sogin M, Morrison H, Huber J, Mark Welch D, Huse S, Neal P, Arrieta JM, Herndl GJ (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103:**

12115-12120.

Theissen U, Hoffmeister M, Grieshaber M, Martin W (2003). Single Eubacterial Origin of Eukaryotic Sulfide:Quinone Oxidoreductase, a Mitochondrial Enzyme Conserved from the Early Evolution of Eukaryotes During Anoxic and Sulfidic Times. *Molecular Biology and Evolution* **20:** 1564-1574.

Wang P, Li T, Hu A, Wei Y, Guo W, Jiao N, Zhang C (2010). Community Structure of Archaea from Deep-Sea Sediments of the South China Sea. *Microbial Ecology* **60:** 796-806.

Weisburg W, Barns S, Pelletier D, Lane D (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology* **173:** 697-703.

Wright JJ, Mewis K, Hanson NW, Konwar KM, Maas KR, Hallam SJ (2013). Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME journal*.

Yancey PH, Gerringer ME, Drazen JC, Rowden AA, Jamieson A (2014). Marine fish may be biochemically constrained from inhabiting the deepest ocean depths. *Proceedings of the National Academy of Sciences* **111:** 4461-4465.

Chapter 5

Concluding Remarks

The main objective of this dissertation was to assess the diversity and metabolic capabilities of microbes present in the deepest portions of the Atlantic and Pacific oceans. The research conducted throughout this dissertation has expanded our current knowledge regarding deep ocean microbial phylogenetic diversity and metabolic functions. The use of single-cell genomics enabled the sequencing of 23 partial single amplified genomes (SAGs). In all cases these organisms signify the deepest representative of their respective groups to be studied to date.

The samples analyzed for these studies were collected from the deepest locations within the Puerto Rico Trench and the Mariana Trench. In the Atlantic Ocean, the Puerto Rico Trench (PRT) is the deepest location. Four single cell genomes were sequenced and analyzed from the PRT samples and compared to closely related surface genomes. Phylogenetic analyses of all four SAGs indicated that they were derived from autochthonous residents of deep-ocean environments, as opposed to microorganisms that were entered the trench benthos as a result of transport from shallow-water settings. Genes present in the SAGs but absent in their comparison genomes revealed novel metabolic capabilities including those associated with nitrogen, sulfur, carbon, and energy acquisition mechanisms. These novel metabolic properties provide them with the capability of utilizing different substrates and pathways to obtain the energy and nutrients needed to sustain their lives. The importance of osmoregulation in the ultra-deep ocean, which may be linked with high-pressure adaptation, is suggested by the finding of aquaporins in seven (30%) of the genomes analyzed. When the SAGs were compared to the available PRT metagenome, evidence for potential trench-specific adaptations was found. Several SAG genes were observed only in a PRT metagenome and not in other

shallower non-trench deep-sea metagenomes. These results illustrate new genomic features that are likely to provide the organisms with tools needed to live in extreme deep ocean environments.

The Challenger Deep, located in the Pacific Ocean's Mariana Trench, is the deepest location of any ocean on earth. Two different candidate phyla from the sediment samples, collected from the Deepsea Challenge Expedition, were sequenced and analyzed. The candidate phylum OD1 had been previously described as exclusively fermentative with very little metabolic diversity and potential. This study expanded the current knowledge of the metabolic potential associated with the OD1 candidate phylum. The different metabolic processes found to be part of the OD1 cells suggests that these organisms are capable of both oxygen and nitrate respiration, complex carbon degradation, and the ability to respond to environmental stress. The Challenger Deep OD1 cells also possess a relatively high abundance of horizontally transferred genes. If the high relative abundance (5.6%, inferred from their SAG numbers) of these organisms is representative of their overall abundance in the Challenger Deep surficial sediment environment in space and time, then they are likely to exert a major influence on this habitat.

The candidate phyla Marinimicrobia was also analyzed from the SAGs recovered from the Challenger Deep sediment samples. For a microbial group that appears to be abundant in many different environments, including in deep and ultra-deep ocean settings like the Puerto Rico Trench and the Mariana Trench (Eloe et al, 2011; Tarn et al, unpublished), the understanding of their metabolic properties at the phylum level and below is remarkably poor. The results presented in this dissertation suggest that as a CP,

the Marinimicrobia are mostly heterotrophic organisms although the possibility of mixotrophy is also present. Many of the genomes possess respiratory and fermentative genomic signatures, which leads to the conclusion that many Marinimicrobia function as facultative anaerobes. Supplementing energy acquisition by the oxidation of sulfur compounds or carbon monoxide may be used, but this seems to be more prevalent in the deep-sea Marinimicrobia than the comparison genomes. The Marinimicrobia genomes also appear to be actively exchanging genetic material, including that involved in the transport, synthesis and recycling of essential cell components such as amino acids. The incorporation of such genes could facilitate growth and survival in the extreme environment of the Challenger Deep. The information gathered from this study provides a greater understanding of the Marinimicrobia phylum, as well as clues to understanding adaptation to ultra-deep ocean conditions.

Based on these discoveries some suggestion can be made as to how to move forward to better understand deep-sea microbes. For example, osmotic regulation and how it may impact high pressure effects in deep sea microbes could be studied in detail by performing aquaporin genetic experiments in the easily cultured moderate piezophiles *Photobacterium profundum SS9* or *Psychromonas CNPT3* (Vezzi *et al*, 2005; Lauro *et al,* 2013). Aquaporin function could be addressed in experiments performed under high-pressure and atmospheric pressure conditions using gene knock out and knock in techniques to delete and overexpress, respectively, the aquaporin genes. Growth could be tracked and water channel activity could be measured to understand the mechanisms of the water flux. These experiments could follow some of the same procedures used Azad and coworkers when assessing the functional characteristics

and hyperosmotic regulation of aquaporin in *Synechocystis* sp. PCC 6803 (Azad *et al*, 2010).

In terms of future culturing efforts, it seems clear from the results of this dissertation that deep-sea microorganisms are able to supplement their energy requirements by oxidizing compounds such as carbon monoxide and hydrogen sulfide, among others. It may be that these compounds are more suitable substrates at depth, and as a result attempts to culture a greater diversity of deep-sea bacteria might benefit by the addition of these energy sources. For some of the SAGs analyzed sulfur compounds seemed to be an important part of their metabolism, so the addition of diverse sulfur species, including elemental sulfur and various sulfide species might promote sulfur oxidation or sulfide reduction.

It would also be very interesting to understand more fully the characteristics shared by the phylogenetic groups examined in this thesis, as well as among them as members of the deep-sea biosphere. I attempted to address this question by comparing genes associated to COG and KEGG categories, but a larger scale whole genome analysis of everything vs. everything could provide a better understanding of the core metabolic properties of each phylum and also what, if any, genomic characteristics are shared across the deep-sea genomes that provide adaptation to hadal ecosystems.

There were also additional trench SAGs that I was unable to fully characterize and warrant further study. For example, a thaumarcheal SAG obtained from the Challenger Deep is surprisingly similar in terms of its metabolic pathways of lipoylation, GCS, urea degradation and ammonia oxidation to the single cell genome analyzed from the Puerto Rico Trench. This makes the Mariana Trench SAG the second

Thaumarchaeota to have been reported to possess many of these metabolic properties. Also there were a number of SAR11 genomes that were found in the Mariana Trench samples, which could have been compared to the PRT SAR11 samples. Some the SAR11 genomes from the Mariana Trench are closely related to the SAR11 clade V microorganism HIMB59, whose phylogenetic association within the SAR11 group in general has been questioned (Viklund *et al*, 2013). So further study of these SAGs could provide a great opportunity to resolve this evolutionary relationship.

This research would have been impossible without the technological advances in the field of single cell genomics (Lasken and McLane, 2014). The availability of high-throughput techniques made it possible to generate a large repository of single cells and subsequently of sequenced genomes for analysis. The partiality of single cell genomes is always inconvenient when trying to understand the metabolic processes that are harbored within a cell, but assembly technologies targeted to single cell genomes have improved the recovery of "almost complete" single cell genomes (Nurk et al, 2013). The improvements in assembly technology address the variable sequence coverage and the high rate of chimeric sequences found in single cell genomes. Also, novel methods for improved understanding of microbial dark matter have resulted from the combination of multiple novel SAGs and treating their sequencing as an man made mini-metagenome. Given that the mini-metagenomes can be created to be of low phylogenetic diversity their assemblies can result in greater sequence coverage (McLean *et al*, 2013). In the case of deep-sea microbes the potential of using mini-metagenomes could be a great help when trying to recover a higher percentage of sequences from environments of low diversity, as was found when looking at single cells from amphipod guts in the Mariana Trench

(mostly *Psychromonas*-like microbes, not discussed within the dissertation). Also the development of low diversity large scale batch cultures at high pressure could provide the option of sequencing whole genomes from organisms of interest without the need of pure cultures. These may prove especially useful for understanding syntrophic microbial communities. Similarly single cells could be captured in gel microdroplets and grown together under high-pressure conditions in which the necessary signals and growth factors required for growth provided by the whole community are present, while still prviding the opportunity to grow pure cultures of colonies within microdroplets for later extraction, amplification and sequencing (Dichosa *et al*, 2014).

From a different angle, improvements the single cell genome assembly area could also be targeted towards improving the combined assemblies of multiple sequenced single cell genome. The possibility of efficiently assembling multiple highly similar single cells into one assembled genomes, while still taking into account the potential for uneven coverage of the amplified genomes, will provide more complete or even totally complete genomes. Another important technological advance that will significantly improve the information that we can gather from single cell genomes is the amplification of single cell cDNA from the single cell mRNA. Having information about what one cell is actively transcribing at a given moment will provide a truer picture of the metabolic function of microbes in hadal environments. Another way to understand more about the active microbial community could come from combining activity assays with single cell genomics. This can be done by sorting single cells that have been fluorescently tag based of their ability to actively metabolize a given substrate (Martinez-Garcia *et al*, 2012). Beyond single cell genomics, metagenomic and metatranscriptomic studies that target the

hadal condition will be needed to better highlight the evolutionary, genetic and regulatory changes required for bacterial and archaeal life in the deepest portions of the world's ocean. Comprehensive analysis of hadal environments will be necessary to understand the ecosystem functioning of the desired system. A combination of 16S rRNA gene surveys, metagenomic, metatrasncriptomic and single cell genomics will be indispensable when trying to uncover significant contributing organisms, their phylogenetic relationships and the metabolic profiles of the active hadal microbial community. In order to run the experiments needed to realize future scientific breakthroughs in this emerging area, large amounts of deep ocean sediment and water samples are needed. The need for robust sampling techniques that provide larger amounts of material will be necessary, as well as samplers that can also maintain in-situ. The development of in-situ filtration systems to sample microbial communities for large-scale molecular work will also provide great insight into the hadal microbial community of microbes in their native environment without the disturbance that may be created by the collection and decompression associated with filtration of large volumes after collection. Although advances in the field of ultra deep-sea microbiology are happening every day, the road to understanding the microbial community of such environments is long, the technological advances and incremental discoveries focusing of on hadal ecosystems will take us closer to understanding the metabolic potential and environmental adaptation of deep-sea microbes.

REFERENCES

Azad AK, Sato R, Ohtani K, Sawa Y, Ishikawa T, Shibata H (2011). Functional characterization and hyperosmotic regulation of aquaporin in< i> Synechocystis</i> sp. PCC 6803. *Plant science* **180:** 375-382.

Dichosa AEK, Daughton AR, Reitenga KG, Fitzsimons MS, Han CS (2014). Capturing and cultivating single bacterial cells in gel microdroplets to obtain near-complete genomes. *Nat Protocols* **9:** 608-621.

Eloe E, Fadrosh D, Novotny M, Allen L, Kim M, Lombardo M Yee-Greenbaum, J, Yooseph, S, Allen, EE, Lasken, R, Williamson, SJ, Bartlett, DH (2011a). Going Deeper: Metagenome of a Hadopelagic Microbial Community. *Plos One* **6**.

Lasken RS, McLean JS (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* **15:** 577-584.

Lauro FM, Stratton, TK, Chastain RA, Ferriera S, Johnson J, Goldberg SMD, Yayanos AA & Bartlett, D. H. (2013). Complete genome sequence of the deep-sea bacterium Psychromonas strain CNPT3. *Genome announcements*, *1*(3), e00304-13.

McLean JS, Lombardo M-J, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, Vyahhi, N, Hall AP, Yang Y, Dupont CL, Ziegler MG, Chitsaz H, Allen AE, Yooseph S, Tesler G, Pevzner PA, Friedman RM, Nealson KH, Venter JC, Lasken RS (2013). Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences* **110:** E2390-E2399.

Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, Xie G, Poulton NJ, Gomez ML, Masland DED, Thompson B, Bellows WK, Ziervogel K, Lo CC, Ahmed S, Gleasner CD, Detter CJ, Stepanauskas R (2012). Capturing Single Cell Genomes of Active Polysaccharide Degraders: An Unexpected Contribution of <italic>Verrucomicrobia</italic>. *PLoS ONE* **7:** e35314.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus, Alla Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. (2013). Assembling Single-Cell Genomes and Mini-Metagenomes From Chimeric MDA Products. *Journal of Computational Biology*. October **20**: 714-737.

Vezzi A, Campanaro S, D'Angelo M, Simonato F, Vitulo N, Lauro F, Cestaro A, Malacrida G, Simionati B, Cannata N, Romualdi C, Bartlett DH, Valle G (2005). Life at depth: Photobacterium profundum genome sequence and expression analysis. *Science* **307:** 1459-1461.

Viklund, J., Martijn, J., Ettema, T. J., & Andersson, S. G. (2013). Comparative and phylogenomic evidence that the alphaproteobacterium HIMB59 is not a member of the oceanic SAR11 clade. *PloS one*, *8*(11), e78858.