Metabolic Diversity within the Globally Abundant Marine Group II Euryarchaea Drives Ecological Patterns

Benjamin J Tully[1,2]

1. Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA
2. Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA, USA

Abstract

Despite their discovery over 25 years ago, the Marine Group II *Euryarchaea* (MGII) have remained a difficult group of organisms to study, lacking cultured isolates and genome references. The MGII have been identified in marine samples from around the world and evidence supports a photoheterotrophic lifestyle combining phototrophy via proteorhodopsins with the remineralization of high molecular weight organic matter. Divided between two Orders, the MGII have distinct ecological patterns that are not understood based on the limited number of available genomes. Here, we present the comparative genomic analysis of 322 MGII genomes, providing the most detailed view of these mesophilic archaea to-date. This analysis identified 17 distinct Family level clades including nine clades that previously lacked reference genomes. The metabolic potential and ecological distribution of the MGII genera revealed distinct roles in the environment, identifying algal-saccharide-degrading coastal genera, protein-degrading oligotrophic surface ocean genera, and mesopelagic genera lacking proteorhodopsins common in all other families. This study redefines the MGII and provides an avenue for understanding the role these organisms play in the cycling of organic matter throughout the water column.

Main text

Since their discovery by DeLong[1] (1992), despite global distribution and representing a significant portion of the microbial plankton in the photic zone, the Marine Group II (MGII) *Euryarchaea* have remained an enigmatic group of organisms in the marine the environment. The MGII have been predominantly identified in the surface oceans[2], account for ~15% of the archaeal cells in the oligotrophic open ocean[3], and shown to increase in abundance in response to phytoplankton blooms[4] comprising up to ~30% of the total microbial community[5]. Research has shown that the MGII correspond with specific genera of phytoplankton[6], during and after blooms[7], and can be associated with particles when samples are size fractionated[8]. Phylogenetic analyses have revealed the presence of two dominant clades of MGII, referred to as MGIIA and MGIIB (the MGIIB have recently been named *Thalassoarchaea*[9]), that respond to different environmental conditions, including temperature and nutrients[10].

To date, the MGII have not been successfully cultured or enriched from the marine environment. Instead our current understanding of the role these organisms play in the environment is derived from interpretations of ecological data (*i.e.*, phytoplankton- and particle-associated) and a limited number of genomic fragments and reconstructed environmental genomes. Collectively, these genomic studies have revealed a number of re-occurring traits common to the MGII, including: proteorhodopsins in MGII sampled from the photic zone[11], genes targeting the degradation of high molecular weight (HMW) organic matter, such as proteins, carbohydrates, and lipids, and subsequent transport of constituent components into the cell[9,12-14], genes representative of particle-attachment[8,12], and genes for the biosynthesis of tetraether lipids[9,15]. Comparatively, the capacity for motility via archaeal flagellum has only been identified in some of the recovered genomes[9,12].

The global prevalence of the MGII and their predicted role in HMW organic matter degradation make them a crucial group of organisms for understanding remineralization in the global ocean. Evidence supports specialization of MGIIA and MGIIB to certain environmental conditions, but the extent of this relationship in the oceans are not understood and cannot be discerned from the available genomic data. The environmental genomes reconstructed from the *Tara* Oceans metagenomic datasets[16-19] provide an avenue for exploring the metabolic variation between the MGIIA and MGIIB, and corresponding metadata collected from the same filter fractions and sampling depths[20,21] can used to understand the ecological conditions that favor each clade. Here, the analysis of 322 non-redundant MGII genomes

identifies the metabolic traits unique to the genomes derived from the MGIIA and MGIIB genomes, providing new context for the ecological roles each clade plays in remineralization of HMW organic matter. Further, the MGIIA and MGIIB can be assigned to 17 Family-level groups, with distinct ecological patterns with respect to sample depth, particle size, temperature, and nutrient concentrations.

Results

Despite their global abundance and active role in the cycling of organic matter, it has been difficult to glean metabolic information from the MGII *Euryarchaea*. As of January 2018, a total of 20 MGII genomes with sufficient quality metrics (>50% complete and <10% contamination) had been reconstructed from environmental metagenomic data and analyzed[9,12,15,22,23]. This number could be supplemented with two single amplified genomes (SAGs) accessed from JGI that were determined to be ~40% complete but possessed 16S rRNA gene sequences. These publicly available genomes were severely skewed towards the MGIIB[15,22,23] (16 genomes) with only six genomes for the MGIIA available[12,15,22]. For the purpose of this study, these 22 previously analyzed genomes are termed the 'Reference Set'. A combined 407 genomes reconstructed from marine environmental metagenomes, originating from four studies utilizing the *Tara* Oceans dataset (designations TMED[16], TOBG[17], UBA[18], and TARA-MAG[19]) and the Red Sea (designated as REDSEA[24]), were identified in publicly available databases. A phylogenetic tree using 16 concatenated ribosomal marker proteins was constructed for the 429 genomes and used to identify genomes originating from the *Tara* Oceans metagenomes with identical branch positions and sample sources (Supplemental Figure 1; Supplemental Table 1). Using completion and contamination metrics, identical genomes were reduced to a single representative, resulting in a dataset of 322 non-redundant MGII genomes (Figure 1). MGIIA and MGIIB formed two distinct branches with a majority of genomes (n = 205) belonging to the MGIIB. The genomes further clustered into 17 distinct clades – 8 MGIIA clades and 9 MGIIB clades. Nine of the clades had no representative from the Reference Set and were composed exclusively of genomes reconstructed from the *Tara* Oceans metagenomic dataset. Based on the extrapolated genome size for these 17 clades, MGIIA genome sizes were significantly larger than MGIIB genomes, on average ~400kbp (Figure 2A; two-sample unequal variance Student's t-test, $p << 0.001$). The two most basal clades of the MGIIB have mean genome sizes similar to that of the MGIIA. In contrast, there was no clear relationship between %G+C content and phylogenetic group; %G+C content of the genomes had a wide range of values (~35%->60%; Supplemental Figure 2). Additionally, several clades had high internal variation of %G+C content. Further splitting clades into 33 subclades, based on the phylogenetic tree and pairwise genome amino acid identity (Supplemental Figures 3 & 4), generated more concise groupings with consistent %G+C values (Figure 2B).

A candidate nomenclature for the MGII based on the reconstructed phylogeny is proposed which incorporates previously proposed names and is further corroborated with details regarding pairwise amino acid identity, metabolic potential, and global abundance patterns. Previous work had proposed that the MGIIB be classified at the Class level under the name *Thalassoarchaea*, in part due to the lack of MGIIA in the marine environment[9]. This has caused some confusion in the literature[25,26] with the name 'Thalassoarchaea' ascribed to all members of the MGII. This research indicates that the MGII represent a Class within the *Euryarchaea*, with the MGIIA and MGIIB representing Order level phylogenetic clades, both of which are present in the marine environment (see below). It is instead proposed here that the name *Thalassoarchaea* be applied to the MGII, with the MGIIA and MGIIB clades reclassified at the Order level with the names *Delongarchiales* and *Valerarchiales*, respectively, to recognize Drs. Edward DeLong and Francisco Rodríguez-Valera for their roles in identifying and studying the ecology of the *Thalassoarchaea*. For assignment at the Family and Genus level, due the propensity of the *Thalassoarchaea* for sunlit environments and consumption of organic matter (see below) akin to the Hobbits from J.R.R. Tolkien's *The Lord of the Rings*, a naming structure that utilizes names associated with towns in the fictional regional known as the Shire for the 17 identified Families and the surnames of Hobbit families for the 33 Genera is proposed (Table 1). Several genomes (n = 35) could not be assigned

at the Family or Genus level and we believe this naming scheme provides an avenue for adding formalized phylogenetic clades in the future.

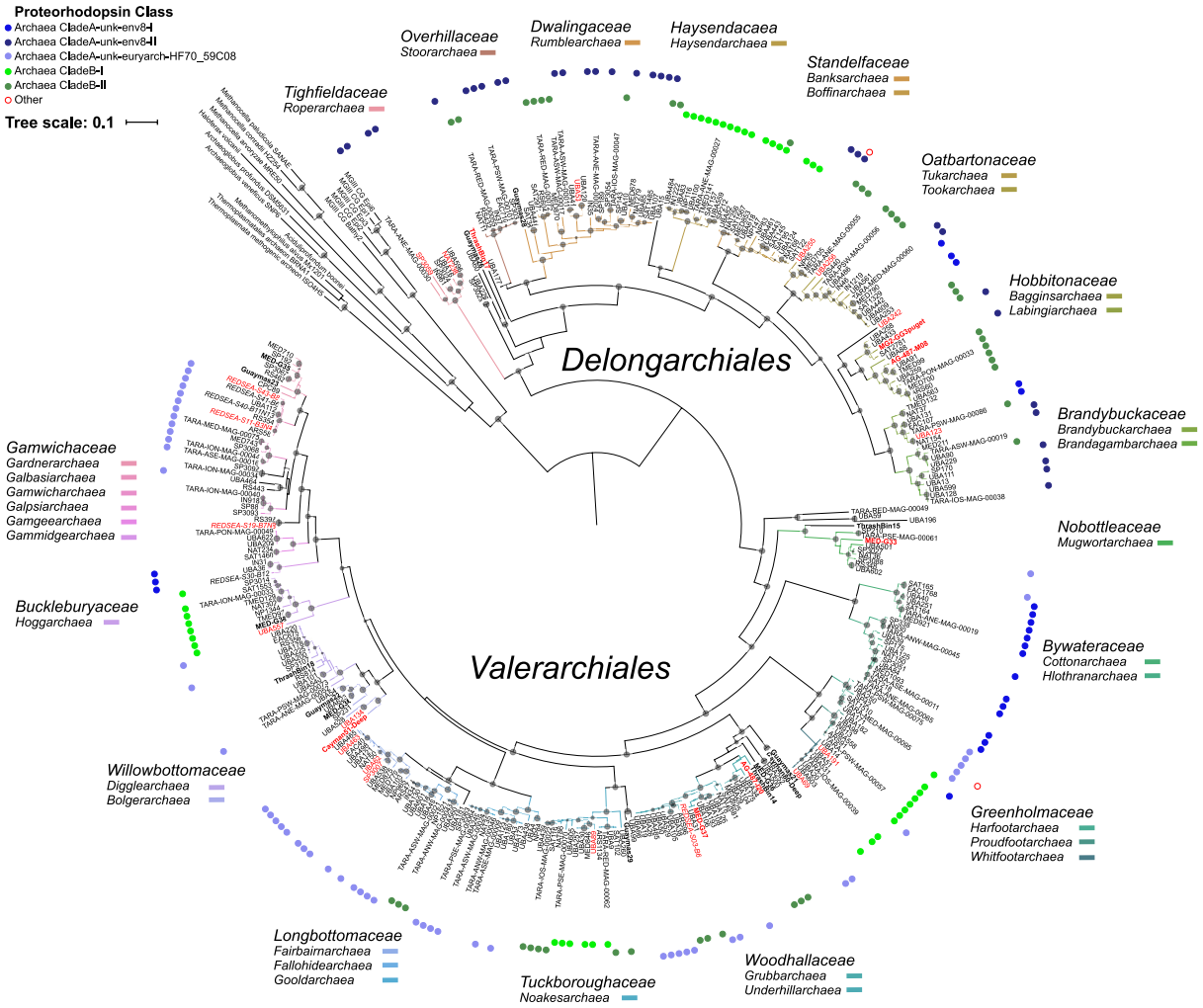


Figure 1 A phylogenomic tree constructed using 16 concatenated ribosomal marker proteins for the Thalassoarchaea. Genomes that represent the 'Reference Set' are in bold. Genomes with an identified 16S rRNA gene sequence are in red. Proposed Family names and the corresponding Genera names are displayed with genomes assigned to a specific genus denoted using the displayed color coded. Bootstrap values are scaled proportionally between 0.75-1. Identified proteorhodopsin classes are denoted with colored circles based on the predicted color of light for which the proteorhodopsin is spectrally tuned.

A subset of the *Thalassoarchaea* genomes had 16S rRNA gene sequence (n = 35) which were used to determine the relationship between previously identified sequence clusters[9,27] and the newly identified families (Supplemental Figure 5). The *Tighfieldaceae* from the *Delongarchiales* and the *Gamwichaceae* and *Nobottleceae* from the *Valerarchiales* were not represented in previously identified *Thalassoarchaea* 16S rRNA gene clusters. Conversely, the previously identified N cluster and clades of the L and O clusters did not have representative environmental genomes, either as a result of missing diversity among the described genomes or due to the fact that not all *Thalassoarchaea* families had a representative 16S rRNA gene present. Some currently defined 16S rRNA clusters corresponded directly to families with genomic representatives; the WHARN cluster to the *Tuckboroughaceae*, the M cluster to the *Oatbartonaceae*, and the K cluster to *Overhillaceae*. The two largest clusters, L from the

Delongarchiales and O from the *Valerarchiales*, were divided at several internal nodes that could be ascribed to two and five of the newly named families, respectively.

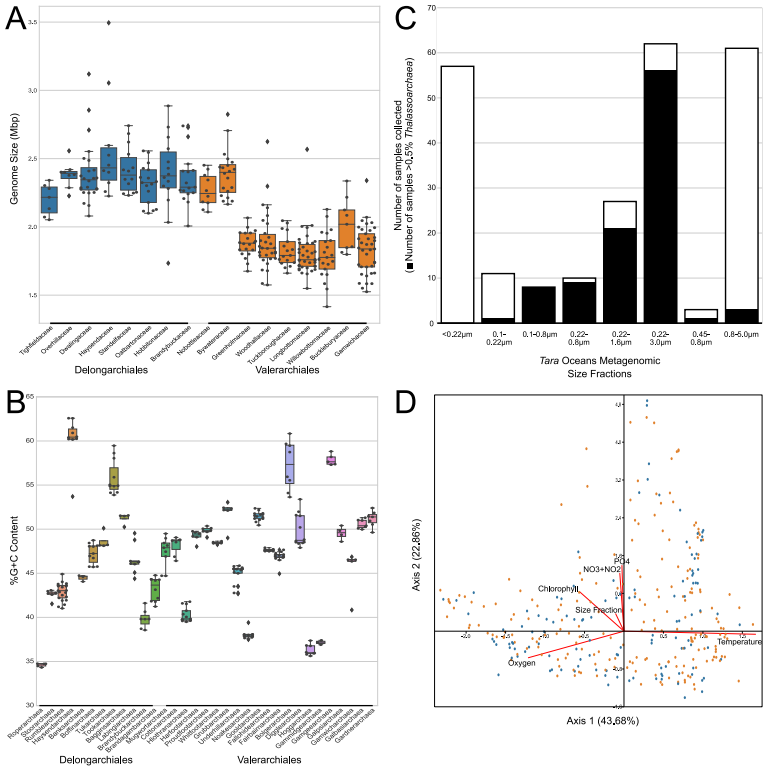


Figure 2 *A. Box plots illustrating the distribution of genome sizes at family level. Delongarchiales and Valerarchiales Family box plots are displayed in blue and orange, respectively. B. Box plots illustrating the distribution of genome GC content at the genus level. Genus box plots utilize the same color scheme as displayed in Figure 1. C. A bar graph where the outline illustrates the number of metagenomic samples available from the Tara Oceans dataset for a given filter fraction and solid filled portion represents the number of samples in that size fraction that recruit ≥0.5% of metagenomic reads to the Thalassoarchaea. D. A canonical correspondence analysis (CCA) based on the RPKM values for the Thalassoarchaea genomes from the high abundance (≥0.5% relative fraction) samples (n = 99). For clarity, the major gradients of the explanatory variables are highlighted in red and amplified 2×. Delongarchiales and Valerarchiales genomes are displayed in blue and orange, respectively.*

***Thalassoarchaea* share an electron transport chain with putative Na^+ pumping components.** There were several shared traits amongst the *Delongarchiales* and the *Valerarchiales*, particularly related to the components of the thalassoarchaeal electron transport chain (ETC). Genomes belonging to both groups had canonical NADH dehydrogenases and succinate dehydrogenases that link electron transport to oxygen as a terminal electron acceptor via low-affinity cytochrome *c* oxidases (Figure 3). As has been noted previously[8], most members of the *Thalassoarchaea* possessed genes encoding a cytochrome *b* and a Rieske iron-sulfur domain protein but lacked the genes for the canonical cytochrome bc_1 complex. Many of the *Thalassoarchaea* families also possessed RnfB, an iron-sulfur protein that can accept electrons from ferredoxin and transfer them to the ETC. The complete Rnf complex is capable of generating a Na^+ gradient through the oxidation of ferredoxin but all members of *Thalassoarchaea* lacked the subunits needed to complete the complex (RnfACDEG). Thus, it was surprising that distributed across all of the families in 240 genomes, the *Thalassoarchaea* possessed an A_1A_O ATP synthase that, based on the presence of specific motifs in the *c* ring protein (AtpK), could be inferred to generate ATP through the pumping of Na^+ ions. All of the genomes had the necessary conserved glutamine and a motif in respective transmembrane helices[28] (Supplemental Figure 6A). The motif in the second helix appears to be diagnostic of the Order a genome belongs to: the *Delongarchiales* contained a LPESxxI motif and the *Valerarchiales* contained a LPETIxL motif. The presence of these motifs does not preclude ATP

synthesis via H^+ pumping[29], though a majority of the experimentally confirmed A_1A_O ATP synthases with these motifs exclusively pump Na^+ ions[28].

***Thalassoarchaea* share the ability to degrade extracellular proteins and fatty acids.** As has been reported previously[9,12-14], a majority of the *Thalassoarchaea* families are poised to exploit HMW organic matter. The families share the potential to degrade and import proteinous material with two extracellular peptidases (sedolisin-like peptidases and carboxypeptidase subfamily M14D) and an oligopeptide transporter present in most of genomes (Figure 3). All of the *Thalassoarchaea* families appear capable of some degree of fatty acid degradation due to the presence of acyl-CoA dehydrogenase and acetyl-CoA C-acetyltransferase, though some of the intermediate steps are missing from all genomes in several families (Figure 3). It is unclear if the incomplete nature of the pathway in these families is the result of uncharacterized family-specific analogs or some degree of metabolic hand-off between different organisms degrading fatty acids. Several other metabolic traits that had been reported in genomes belonging to either the *Delongarchiales* or *Valerarchiales* are also part of the thalassoarchaeal core genome[9,15], including the capacity for the assimilatory reduction of sulfite to sulfide, the transport of phosphonates, flotillin-like proteins, which may have a role in cell adhesion, and geranylgeranylglyceryl phosphate (GGGP) synthase, a key gene for tetraether lipid biosynthesis (Figure 3).

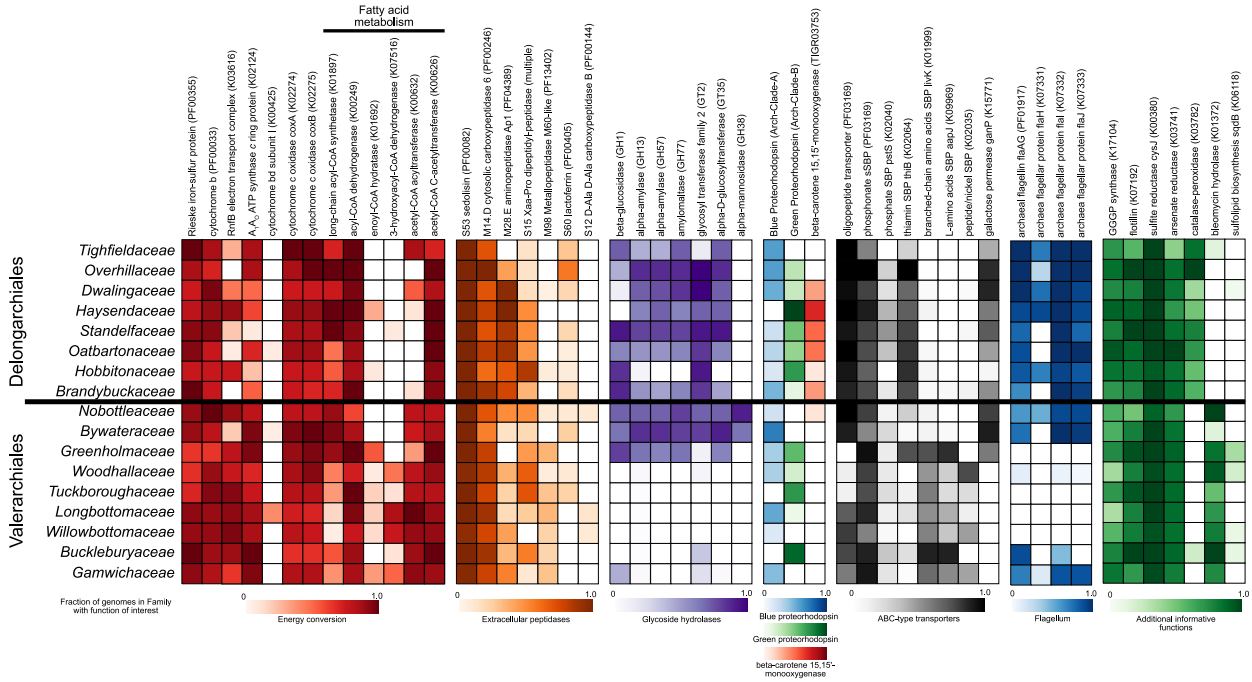


Figure 3. Heatmap of the occurrence of various functions of interest in the respective Thalassoarchaea families. Heatmaps are scaled from 0 to 1, where 1 represents that all genomes within the designated Family possess the function of interest.

Putative proteorhodopsins differentiate members of the *Delongarchiales* and *Valerarchiales*. While components of the ETC and HMW degradation were present in all thalassoarchaeal families, there were several traits that either lacked a phylogenetic signature or differentiated the *Delongarchiales* and the *Valerarchiales*. As has been noted previously[12] and confirmed with this collection of genomes, all of the *Thalassoarchaea* families possess genes encoding light-sensing rhodopsins and, based on the amino acids at positions 97 (aspartate) and 108 (lysine/glutamic acid) in the rhodopsin sequences, are predicted to function as proteorhodopsins capable of establishing H^+ gradients (Supplemental Figure 6B). Phylogenetically, these proteorhodopsins (PRs) cluster in established clades[30] Archaea Clade A (Clade-A) and Archaea Clade B (Clade-B) and based on the amino acid in position 105 (glutamine/methionine), spectral tuning prediction indicates sensitivity to blue and green light, respectively (Supplemental Figure

6B). Five families exclusively possess Clade-A, three families exclusively possess Clade-B, and nine families have genomes that possess either of the two PRs. Only two genomes possessed both PR clades.

The *Bolgerarchaea* (Family *Willowbottomaceae*), which contains a number of thalassoarchaeal genomes reconstructed from the deep-sea, do not possess PRs (Figure 1). The lack of PRs in deep-sea *Thalassoarchaea* is consistent across the tree, with deep-sea reconstructed genomes not present in the *Bolgerarchaea* tending to represent the most basal branching members of other families (*e.g.*, genome Guaymas21 within the Family *Woodhallaceae*). Three genera (*Gamgeearchaea*, *Galpsiarchaea*, and *Gardnerarchaea*) within the *Gamwichaceae* also lack identifiable PRs. Proteorhodopsins from Clade-A fall into three distinct phylogenetic groups associated with the clades unk-env8 (CladeA-unk-env8-I and -II) and unk-euryarch-HF70_59C08 identified in the MICrhoDE database, while Clade-B has two distinct groups (Clade-B-I and -II) (Supplemental Figure 7). The *Delongarchiales* possessed all of the PR groups, except unk-euryarch-HF70_59C08 and slightly favor the green light tuned PRs (54% of PR containing genomes), while the *Valerarchiales* do not utilize the CladeA-unk-env8-II group and favor blue light tuned PRs (64% of PR containing genomes). Additionally, several families and genera possessed exclusively one of the PR clades (Figure 1). Despite the requirement of the chromophore retinal for the functioning of PR, a majority of the *Thalassoarchaea* lacked an annotation for beta-carotene 15,15'-monooxygenase (Figure 3), essential for the last cleavage step needed to activate retinal. Two of the eight families from the *Delongarchiales* and all but one of the families from the *Valerarchiales* lacked this crucial functional step.

The degradation of extracellular peptidases and algal oligosaccharides differentiate members of the *Delongarchiales* and *Valerarchiales*. While the *Thalassoarchaea* shared several functionalities with a role in the degradation of HMW organic matter, there was a greater diversity of functionality in specific orders and families (Figure 3). There were five additional classes of extracellular peptidases (aminopeptidases subfamily M28E, dipeptidyl-peptidase, M60-like metallopeptidase, lactoferrin-like, and carboxypeptidase B) common (and 16 extracellular peptidases with infrequent occurrence; Supplemental Table 2) amongst the genomes. The collective suite of peptidases within a genome dictate the potential types of proteinous material that be processed by an organism. Three of the five extracellular peptidase classes were distributed across both the *Delongarchiales* and *Valerarchiales*, while the M60-like metallopeptidase and carboxypeptidase B, were present almost exclusively amongst the *Valerarchiales*. Despite sharing many of the putative protein degrading functions, families from the *Valerarchiales*, except for *Nobottleaceae* and *Bywateraceae*, possess the substrate-binding proteins for ATP-binding cassette (ABC) type transporters for three additional amino acid and peptide transporters (branched-chain amino acids, L-amino acids, and peptide/nickel), while the *Delongarchiales* only have the previously noted oligopeptide transporter (Figure 3).

Beyond the degradation of proteins and fatty acids, there is evidence to suggest that *Thalassoarchaea* have a role in the degradation of carbohydrate HMW organic matter[31]. Interestingly, glycoside hydrolases with functionality for the degradation of algal oligosaccharides, including pectin, starch, and glycogen, are found exclusively amongst the *Delongarchiales* and the most basal families of the *Valerarchiales*, the *Nobottleaceae* and *Bywateraceae* (Figure 3). These same clades also possess an annotated galactose permease subunit for an ABC-type transporter. Further, *Nobottleaceae* and *Bywateraceae* also possess a glycoside hydrolase that could possibly play a role in mannosylglycerate degradation, an osmolyte found in red algae[32].

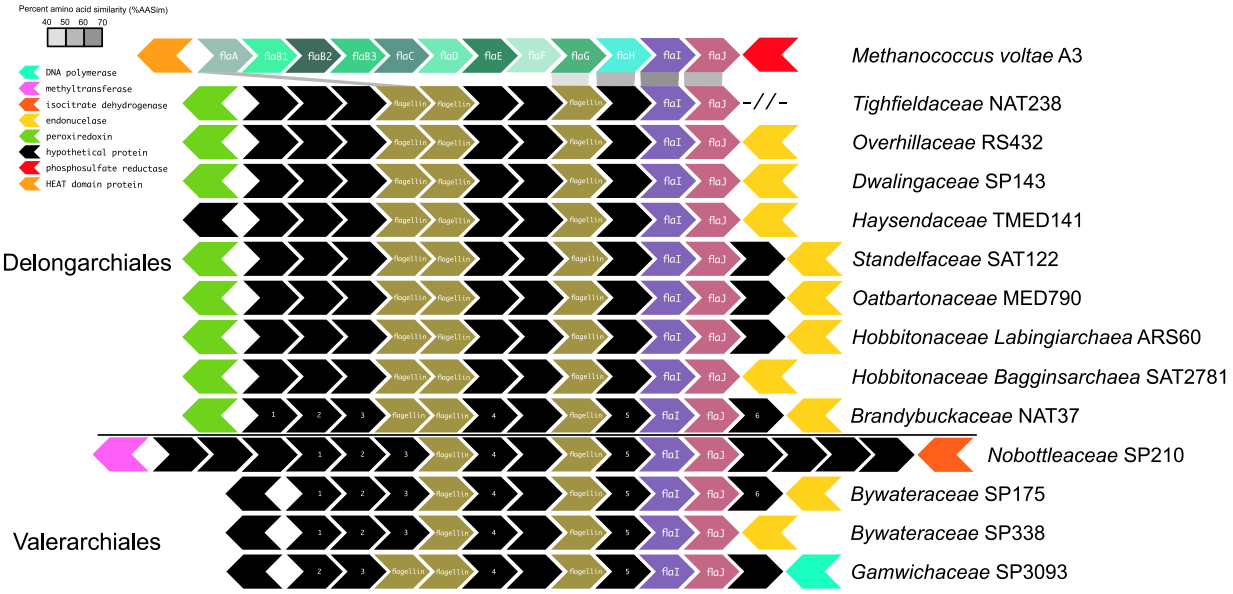


Figure 4. A stylized view of the putative motility genes present in the Thalassoarchaea. The longest contig for each Family (or Genus) is shown. Hypothetical proteins lacking KEGG annotations or matches to queried HMMs are in black. All hypothetical proteins in a column had significant BLAST matches to their neighbors, except as noted by the numbers in the transition from the Delongarchiales to the Valerarchiales. Flagellins noted in the gold segment were detected using the archaeal flagellin PFAM (PF01917). Proteins immediately upstream and downstream are colored based on predicted function. Significant BLAST matches between Methanococcus voltae A3 and the Tighfieldaceae genome are noted.

Motility is a trait common to the *Delongarchiales*. Previous research has shown evidence for and against the putative capacity for motility amongst the *Thalassoarchaea*[9,12]. The thalassoarchaeal genomes lacked annotations or homology for most of the canonical archaeal flagellum operon (Figure 4). However, genomes from all of the *Delongarchiales* families, *Nobottleaceae*, *Bywateraceae*, and *Gamwichaceae* possessed proteins annotated as subunits from the canonical operon (FlaAGHIJ). A comparison of the identified subunits from a representative of the *Roperachaea* to *Methanococcus voltae* A3 revealed 40-70% amino acid similarity between putative orthologs. These subunits were syntenic in a region that contained an additional 1-3 identifiable flagellins and several orthologous proteins lacking annotations. All of the predicted proteins in this region could be identified by similarity between representatives of each family. The structure of the region, including the predicted proteins immediately up- and downstream of the region, appeared to be mostly conserved amongst the *Delongarchiales*, while some variation in gene content could be observed amongst the clades from the *Valerarchiales*.

For several other functions ascribed to the *Thalassoarchaea* as a whole[9], there are distinct distributions amongst the orders, including the presence of a catalase-peroxidase amongst the *Delongarchiales* and a bleomycin hydrolase amongst the *Valerarchiales* (Figure 3). Further, several other predicted metabolic functions appear to be specific to only a subset of families and may have a role in niche differentiation amongst the thalassoarchaeal families, including cytochrome *bd* (a high-affinity oxygen cytochrome responsible for microaerobic respiration), a phosphate substrate-binding subunit for an ABC-type transporter, and UDP-sulfoquinovose synthase, a key gene for the biosynthesis of sulfolipids (Figure 3).

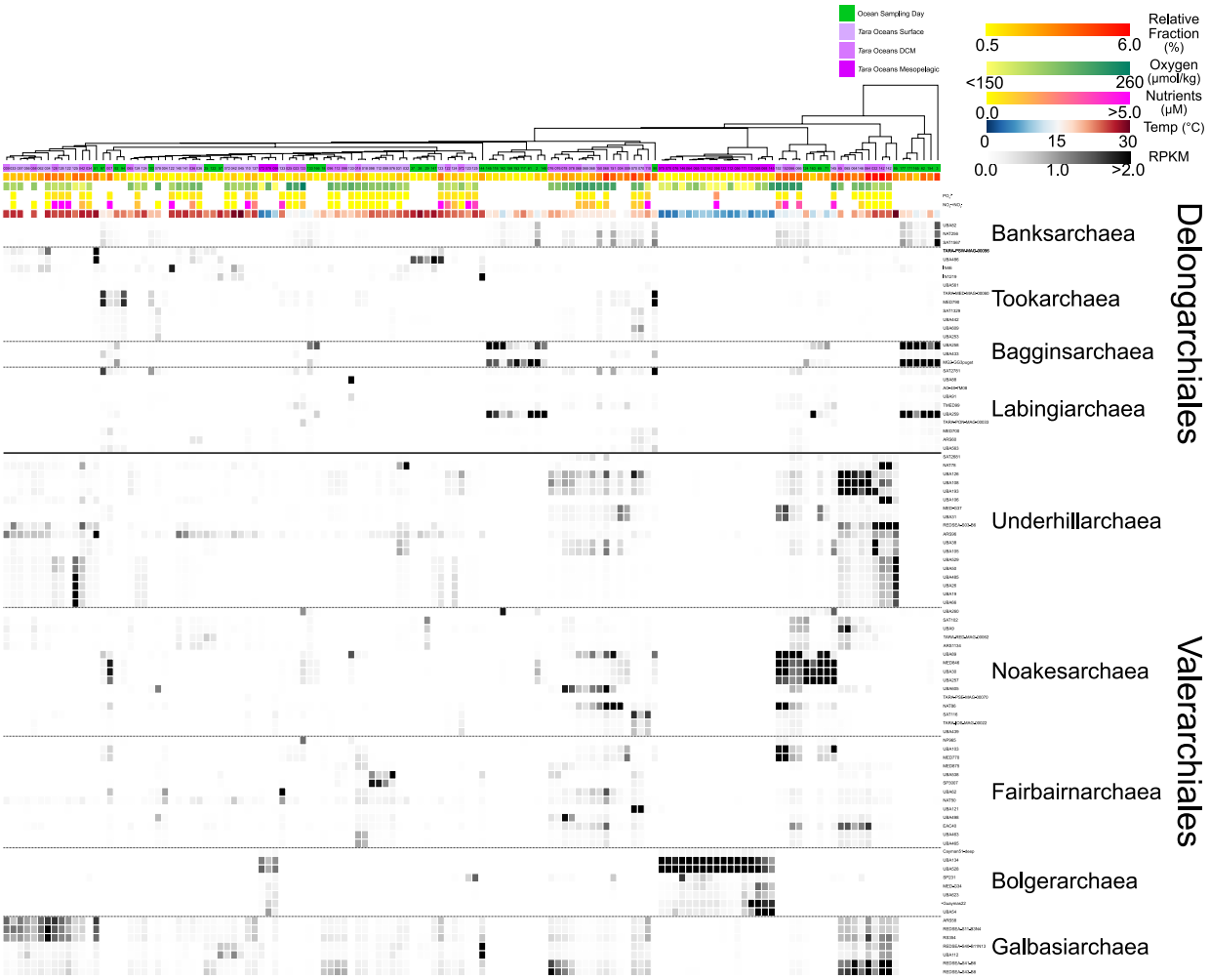


Figure 5. A heatmap displaying the RPKM values for a subset of Thalassoarchaea *genomes discussed in the manuscript in high abundance samples (≥0.5% relative fraction). RPKM values are scaled from 0-2 with values ≥2 in black (median, 0.001; maximum, 31.54). Samples are hierarchically clustered based on all* Thalassoarchaea *RPKM values and the sample source is displayed as either green (OSD) or purple (*Tara *Oceans). Numbers displayed for samples correspond to sample/station ID. The available environmental parameters are presented as colored heatmaps (missing parameters are displayed as white). A heatmap displaying the RPKM values for all thalassoarchaeal genomes is available in Supplemental Figure 9.*

Genera from the *Thalassoarchaea* inhabit distinct marine niches. Using a comprehensive set of *Tara* Oceans metagenomic datasets from across the globe[21,33], that included all of the size fractions for which DNA was collected (viral, 'bacterial', and eukaryotic), it was possible to explore where specific thalassoarchaeal groups were dominant. The *Thalassoarchaea* were rarely found to be abundant (>0.5% relative abundance; mean, 2.13%; maximum, 6.07%) in samples for size fractions <0.22µm or >0.8µm, with almost all abundant samples occurring in the 'bacterial' size fractions (0.1-3.0µm; Figure 2C). Globally, the *Thalassoarchaea* were abundant at all *Tara* Oceans stations with a 'bacterial' size fraction (n = 47), except for at four stations (Supplemental Figure 8). There were no *Tara* Oceans metagenomic samples collected from size fractions >5µm. Examining the most abundant thalassoarchaeal genomes reveals that the *Valerarchiales* tend to be the dominant groups in oceanic samples (Figure 5; Supplemental Figure 9), specifically the *Underhillarchaea*, *Noakesarchaea*, and *Galbasiarchaea*. The *Bolgerarchaea* are only dominant in mesopelagic samples, predominantly to the exclusion of all other genomes, except for some basal groups containing genomes from deep-sea samples (Supplemental Figure 9).

In trying to understand how the environmental parameters may impact the distribution of the *Thalassoarchaea*, genome abundance metrics were subjected to a canonical correspondence analysis for samples with high abundance of *Thalassoarchaea*. The major drivers of thalassoarchaeal occurrence were oxygen, temperature, and nutrients (phosphate and nitrate [nitrate refers to the combined measurement of nitrate + nitrite]), however these parameters did not differentiate the two Orders. Conversely, when the *Tara* Oceans samples were clustered based on the thalassoarchaeal genome abundance metrics, there were several distinct groups that had unifying physical properties (Figure 5; Supplemental Figure 9). All but three of the mesopelagic samples clustered in a cohesive group with the *Bolgerarchaea* as the most abundant organisms in those samples. The *Noakesarchaea* (Family *Tuckboroughaceae*) were abundant in samples with moderate temperature (14-15°C), high oxygen (235-42 µmol/kg), and high nitrate (2-4µM). While *Galbasiarchaea* are dominant in the tropical samples with high temperature (24-27°C), moderate oxygen (160-90 µmol/kg), and high nitrate (>5µM). The *Galbasiarchaea* were present along with the *Underhillarchea* in high temperature samples (24-26°C), moderate oxygen (180-90 µmol/kg), and low phosphate and nitrate (<0.1µM).

The abundance of the *Delongarchiales* in open ocean samples was limited. In an effort to identify samples where the Order may be abundant and based on previous studies, 118 'prokaryotic' metagenomes from coastal (<10km) Ocean Sampling Day[34] 2014 (OSD) samples were assessed for the presence of the thalassoarchaeal genomes (Figure 5; Supplemental Figure 9). These samples were collected using a unified method that captured whole seawater >0.22µm and measured a limited number of physical properties, generally, temperature, salinity, distance to the coast, and depth (0-5m). Unlike the ubiquitous nature of *Thalassoarchaea* in the 'bacterial' *Tara* Oceans fractions, only about a third of the samples (n = 37) from OSD had high thalassoarchaeal abundance. These samples almost exclusively recruited to the *Delongarchiales*, dominated by the *Banksarchaea*, *Bagginsarchaea*, *Labingiarchaea*, and *Tookarchaea*. Unlike the *Tara* samples, where temperature played a role in determining the dominant thalassoarchaeal genera, OSD samples that cluster together have a much wider range of temperatures (*e.g.*, 14-20°C and 11-21°C), suggesting that temperature plays a less important role in structuring *Thalassoarchaea* abundance/occurrence in these samples. Determining the physical parameters that do correlate with thalassoarchaeal abundance was not possible as OSD samples had fewer measured physical properties compared to *Tara* Oceans samples.

Discussion

The details in phylogeny, metabolism, and ecology provided by the increased resolution of *Thalassoarchaea* genomes collected for this study redefines what is understood about this globally dominant euryarchaeal Class. Previous phylogenetic diversity contained within reconstructed genomes and genomic fragments failed to capture at least nine newly defined Family-level clades. This collection of 322 genomes allows for a precise understanding of the metabolic potential present in the *Thalassoarchaea*, including the metabolic and ecological differentiation of the *Delongarchiales* and *Valerarchiales*.

Core components of the proposed metabolism for the *Thalassoarchaea* remain, including an obligate aerobic heterotrophic-lifestyle oriented around the remineralization of proteins and lipids that compose HMW organic matter with the capacity to harness solar energy through proteorhodopsins. The possibility that thalassoarchaeal A_1A_O ATP synthases can exploit a sodium motive force, as well as a proton motive force, opens an avenue for energy conversion that differs from most marine bacteria and archaea. How this ETC would function *in situ* is unclear but may be linked to the only identifiable component of the Rnf sodium translocating complex, RnfB. It may be that the *Thalassoarchaea* utilize both H^+ and Na^+, similar to *Methanosarcinales* under marine conditions[29], and that different elements of the *Thalassoarchaea* ETC perform these translocations. Further investigations in to the functionality of thalassoarchaeal proteorhodopsins and noncanonical cytochromes may resolve how this ETC differs from other marine microorganisms.

While the degradation of proteins and fatty acids appears to be a staple of thalassoarchaeal heterotrophy, the often reported role in carbohydrate degradation, as established by the first

Thalassoarchaea genome[12], appears to be limited to the *Delongarchiales* and the two most basal families of the *Valerarchiales*. The specificity of the annotated glycoside hydrolases, implies that these members of the *Thalassoarchaea* are exploiting algal derived substrates. However, the most abundant thalassoarchaeal genera in the open ocean lack the capacity to degrade these algal compounds. Assigning environmental 16S rRNA gene sequences to specific thalassoarchaeal genera will be important in shaping how past and future research interprets the potential function of *Thalassoarchaea* sequences in a sample.

The overlap of the different euryarchaeal proteorhodopsin clades, especially in regard to blue and green light spectral tuning, between the two Orders highlights the adaptation of certain groups to localized conditions but may also indicate a larger trend towards the type of light wavelengths available in a particular niche. The mesopelagic dominant *Bolgerarchaea* and other deep-sea *Thalassoarchaea* all lack proteorhodopsins but maintain similar heterotrophic capacity, providing evidence for proteorhodopsin functionality as an indicator of localized adaptation. The putative motility operon is almost exclusively linked to families with the metabolic potential to degrade algal-derived carbohydrates. This relationship may indicate that members of the *Delongarchiales*, *Nobottleaceae*, and *Bywateraceae* use motility to remain in the proximity of algal-derived HMW organic matter sources, while the remaining families in the *Valerarchiales* exploit proteinous HMW without active movement between particles.

The *Thalassoarchaea* represent a globally persistent group of organisms with a role in organic matter remineralization with two Orders specialized for distinct niches. The dominance of the *Valerarchiales* in oligotrophic open ocean environments and not coastal systems may be linked to adaptations such as smaller genomes, in part driven by the loss of metabolic potential for exploiting algal oligosaccharides and motility. There are several distinct ecological patterns of *Valerarchiales* abundance that need to be explored further and determine how the patterns are related to metabolic diversity. For example, the *Galbasiarchaea* and *Underhillarchaea* occur in *Tara* Oceans samples with similar ranges in temperatures and oxygen concentrations, but *Underhillarchaea* are less abundant in sample with high nitrate concentrations. A similar divide also occurs for individual genomes within the *Galbasiarchaea*. Future examination into the mechanisms for nutrient scavenging and susceptibility to toxicity may prove insightful for determining *Valerarchiales* ecological distributions.

The dominance of the *Delongarchiales* in coastal samples appears to be tied to physical parameters other than temperature. *Thalassoarchaea* have previously been identified in filter fractions greater than 3µm and were hypothesized to have been attached to large plankton[8]. It is possible that *Delongarchiales* are more abundant globally in these size fractions, but the lack of metagenomes from >5µm from *Tara* Oceans makes this difficult to assess. Ultimately, large-scale analysis of thalassoarchaeal genomic potential across 17 newly-defined Families allows for the reinterpretation of the role these organisms play in the cycling of HMW organic matter in the environment and opens new avenues for future research.

Table 1 Nomeclature for the proposed Thalassoarcahea Class

Order	Family	Genus
Delongarchiales	Tighfieldaceae	Roperarchaea
	Overhillaceae	Stoorarchaea
	Dwalingaceae	Rumblearchaea
	Haysendaceae	Haysendarchaea
	Standelfaceae	Banksarchaea
		Boffinarchaea
	Oatbartonaceae	Tukarchaea
		Tookarchaea
	Hobbitonaceae	Bagginsarchaea

		Labingiarchaea
	Brandybuckaceae	Brandybuckarchaea
		Brandagambarchaea
Valerarchiales	Nobottleaceae	Mugwortarchaea
	Bywateraceae	Cottonarchaea
		Hlothranarchaea
	Greenholmaceae	Harfootarchaea
		Proudfootarchaea
		Whitfootarchaea
	Woodhallaceae	Grubbarchaea
		Underhillarchaea
	Tuckboroughaceae	Noakesarchaea
	Longbottomaceae	Gooldarchaea
		Fallohidearchaea
		Fairbairnarchaea
	Willowbottomaceae	Bolgerarchaea
		Digglearchaea
	Buckleburyaceae	Hoggarchaea
	Gamwichaceae	Gammidgearchaea
		Gamgeearchaea
		Galpsiarchaea
		Gamwicharchaea
		Galbasiarchaea
		Gardnerarchaea

Methods

Genome Selection and Phylogenetic Assessment

MGII genomes that were publicly available prior to January 1, 2018[12,15,22-24] were collected from NCBI[35] and IMG[36] and were assessed using CheckM[37] to determine the approximate completeness and degree of a contaminating sequences (Supplemental Table 1). A 'Reference Set' of genomes that were >50% complete and <5% contaminated were included in downstream analysis, with the exception of two single-amplified genomes which were ~40% complete but possessed an annotated 16S rRNA gene sequence. Genomes with predicted phylogenetic placement within the MGII that were derived from the *Tara* Oceans metagenomic datasets[16,17,19,38] were collected and assessed with CheckM (as above). Genomes originating from Tully *et al*. (2017, 2018) that had >5% predicted contamination were refined as described in Graham *et al*.[39] (2018). Briefly, high contamination genomes originally binned using BinSanity[40] (v.0.2.6.2) had their sequences pooled with contigs from the same regional dataset (see Tully *et al*. 2018) and were binned based on read coverage and DNA composition data using CONCOCT[41] (v.0.4.1). All new CONCOT bins containing sequences previously binned together with BinSanity were visualized in Anvi'o[42] (v.3) (anvi-profile) and manually refined to reduce the degree of contamination.

Predicted protein sequences from NCBI were used when possible, while genomes lacking formalized coding DNA sequence (CDS) prediction had proteins sequences predicted using Prodigal[35] (v.2.6.3). The predicted proteins sequences for each genome were searched (HMMER[43] v.3.1b2; hmmsearch -E 1E-5) using HMM models representing the 16 predominantly syntenic ribosomal proteins identified in Hug *et al*.[44] (2016) (Supplemental Data 1). All proteins with a match to a ribosomal protein model were aligned using MUSCLE[45] (v.3.8.31; -maxiters 8) and automatically trimmed using trimAL[39] (v.1.2rev59; -automated1). All 16 alignments were concatenated and a phylogenetic tree was constructed using FastTree[40] (v.2.1.10; -gamma -lg). All described phylogenetic trees were

visualized using the Interactive Tree of Life[46]. The phylogenetic tree was used to manually identify genomes derived from the *Tara* Oceans metagenomic datasets (TMED, TOBG, UBA, and TARA) that were phylogenetically identical and originated from the same samples (Supplemental Table 1; Supplemental Data 2). Completion and contamination statistics for identical genomes were compared and the genome with superior values was retained for further analysis. Duplicate genomes were removed from the concatenated alignment and a phylogenetic tree of the non-redundant genome dataset was generated using FastTree (as above; Supplemental Data 3). Pairwise amino acid identity (AAI) was calculated for the genomes from the two major clades (MGIIA and MGIIB) using CompareM (https://github.com/dparks1134/CompareM; v.0.0.23; aai_wf defaults; Supplemental Figure 3 and 4; Supplemental Data 4). Based on the phylogenetic tree and corresponding AAI values a nomenclature to describe the MGII *Euryarchaea* was created.

Genomes originating from environmental metagenomic samples[16-19,24] were assessed for the presence of the 16S rRNA gene using RNAmmer[42] (v.1.2; -S arch -m ssu). Identified sequences were combined with 16S rRNA gene sequences representing the available various reference genomes[12,15,22,23] and previously established clusters[9] (MGIIA clusters K, L, M; MGIIB clusters O, N, WHARN). As above, sequences were aligned using MUSCLE, automatically trimmed using trimAL, and used to construct a phylogenetic tree using FastTree (-nt -gtr). When possible, the previously defined 16S rRNA gene clusters were classified based on the proposed nomenclature, including splitting previous 'monophyletic' clusters (Supplemental Data 4 and 5).

Functional Prediction

A uniform function annotation was applied to all predicted proteins for the non-redundant genomes. Proteins were annotated with the KEGG database[43] using GhostKOALA[44] ('genus_prokaryotes+family_eukaryotes'; accessed December 1, 2017). Extracellular peptidases (enzymes predicted to degrade proteins) were identified with matches (hmmsearch -T 75) to PFAM HMM models[47] corresponding to MEROPS peptidase families[48] (Supplemental Table 3; Supplemental Data 7) that were predicted to have "extracellular" or "outer membrane" localization by PSortb[47] (v.3; -a) or an "unknown" localization with predicted translocation signal peptides by SignalP[49] (v.4.1; -t gram+). Carbohydrate-active enzymes (CAZy)[50] were identified (hmmsearch -T 75) using HMM models from dbCAN[51] (v.6). Functions of interest were predominantly identified based on the corresponding KEGG Orthology (KO) entry and GhostKOALA predictions. Specific functions of interest without a KO entry were searched using HMM models (hmmsearch -T 75) obtained from PFAM and TIGRFAM[52] (v.15.0).

Predicted proteins of each genome were screened for matches to the rhodopsin PFAM model (PF01036; hmmsearch -T 75; Supplemental Data 8). In order to identify putative proteorhodopsins, sequences matching the rhodopsin HMM model were processed using the Galaxy-MICrhoDE workflow implemented on the Galaxy web server (http://usegalaxy.org) to assign rhodopsins to the MICrhoDE database[53]. The alignment generated from the workflow was manually trimmed to a 96 amino acid region conserved across all sequences, re-aligned using MUSCLE and used to construct a phylogenetic tree with FastTree (as above; Supplemental Data 9). The rhodopsins were predominantly assigned to three clades based on the phylogenetic relationships with other MICrhoDE sequences, unk-euryarch-HF70-59C08, unk-env8, and one unassigned clade. Two rhodopsins were assigned to additional clades, MICrhoDE clade IV-Proteo3-HF10_19P19 and a unassigned clade. Based on Pinhassi *et al.* (2016), unk-euryarch-HF70-59C08 and unk-env8 are also known as Archaea Clade-A and the unassigned clade belongs to Archaea Clade-B. A more detailed phylogenetic tree was construct (as above) using only sequences from MGII (Supplemental Figure 7). The MGII rhodopsin sequences were aligned using MUSCLE and were assessed for specific amino acids present at positions 97 and 108 to determine putative function and position 105 to determine putative spectral tuning (Supplemental Figure 6B).

The operon putatively encoding an archaeal flagellum was identified based on the presence of co-localized the flagellar proteins FlaHIJ (K07331-3) and archaeal flagellins (PF01917). All genomes with possible co-localization of these proteins were identified (Supplemental Table 4). Putative operons from non-redundant TOBG genomes were visualized by subclade using the progressiveMauve aligner[54] (v.2.3.1; default) and longest contig containing the operon was selected to represent that subclade (Supplemental Data 10). Each representative was the compared to its phylogenetic neighbor using BLASTP[55] (v.2.2.30+; parameters) to identify orthologs.

MGII Core Genome Analysis

A pangenomic analysis was performed for the genomes belonging to *Delongarchiales* and *Valerarchiales* using the Anvi'o pangenome workflow[56] (v.3). The pangenome analysis was executed on *Delongarchiales* and *Valerarchiales* separately, where genomes from each Genus within in a Family were combined to generate the necessary inputs. Thus, *Delongarchiales* had eight and *Valerarchiales* had nine inputs representing the various Families, where each Family input was composed of all the underlying genomes. The pangenomic analysis within

Anvi'o used the default parameters for minbit[57] (--minbit 0.5) and MCL[58] (--mcl-inflation 2) to generate protein clusters (PCs). Results were visualized in Anvi'o (anvi-display-pan) with the cladogram displayed using gene frequencies. PCs present in all Families or within in a majority of Families (*e.g.*, a subset of PCs present in all *Delongarchiales* subclades except *Roperarchaea*) were identified and the underlying protein sequences were extracted (anvi-summarize).

PCs were determined to represent a function of the *Delongarchiales* or *Valerarchiales* core genome if it contained a number of proteins greater than 70% (i.e., the average completeness of all *Thalassoarchaea* genomes) of the genomes in the clade (*Delongarchiales*, PCs with >78 proteins; *Valerarchiales*, PCs with >141 proteins). Adjustments were made for PCs that were missing from a single Genera (e.g., *Delongarchiales* without *Roperarchaea*, PCs with >73 PCs). Proteins from all core PCs were submitted to GhostKOALA[44] ('genus_prokaryotes+family_eukaryotes'; accessed February 2, 2018) for annotation. The number of proteins assigned to a PC were manually compared to the number of proteins within the PC with a predicted KEGG annotation. PCs where a majority of proteins had the same KEGG assignment were ascribed that putative function. PCs that did not meet this threshold were considered not to have an annotation. PCs with multiple KEGG assignments were ascribed a KEGG function if one predicted function reached the majority threshold, especially if all assignments had similar predicted functions (*e.g.*, multiple ABC-type transporter ATP-binding proteins). The KEGG annotations from *Delongarchiales* were compared to *Valerarchiales* and overlapping functions were determined to be core components of the *Thalassoarchaea* pangenome. KEGG annotations distinct to each Order were determined be to core components of each Order's pangenome (Supplemental Table 5).

MGII Relative Fraction and Environmental Correlations

The non-redundant set of MGII genomes were used to recruit sequences from environmental metagenomic libraries, specifically 238 samples from *Tara* Oceans representing 62 stations and 118 samples from Ocean Sampling Day (OSD) 2014[59] (Supplemental Table 6). Metagenomic sequences were recruited using Bowtie2[58] (v.2.2.5; --no-unal). Resulting SAM files were sorted and converted to BAM files using SAMtools[60] (v.1.5; view; sort). featureCounts[60] (v.1.5.0-p2; default parameters) implemented through Binsanity-profile[40] (v.0.2.6.4; default parameters) was used to generate read counts for each contig from the sorted BAM files (Supplemental Data 11). Read counts were used to calculate the relative fraction of each *Thalassoarchaea* genome in all metagenomic samples (reads recruited to a genome ÷ total reads in metagenomic sample) and reads per kbp of each genome per Mbp of each metagenomic sample (RPKM; (reads recruited to a genome ÷ (length of genome in bp ÷ 1000)) ÷ (total bp in metagenome ÷ 1000000)) (Supplemental Data 12). Samples were divided into high (≥0.5% MGII recruitment) and low relative fraction samples (<0.5% MGII recruitment). Based on these designations, RPKM values for *Thalassoarchaea* genomes from *Tara* Oceans samples with high relative fraction with sufficient metadata (filter size fraction, depth, temperature, and oxygen, chlorophyll, phosphate, and nitrate [measured as nitrate + nitrite]), were used in a canonical correspondence analysis (CCA) in Past3[61] (v.3.20). Due the correlation of depth with a number of factors, temperature, chlorophyll, phosphate, and nitrate, depth was removed from the final CCA (data not shown). OSD samples consistently only collected temperature, distance from the coast, and salinity. RPKM values for *Thalassoarchaea* genomes from high relative fraction samples were clustered using Ward hierarchical clustering with Euclidean distances implemented with SciPy (http://www.scipy.org; v.1.0.0) and visualized with seaborn (http://seaborn.pydata.org; v.0.8.1). Hierarchical clustering was performed for the *Tara* Ocean samples, the OSD samples, and both datasets combined.

Data Availability

The genomes used in this study are publicly available, except for a subset of the 'Reference Set' from Li *et al*. (2015) which were provided by personal communication, and reference IDs are available in Supplemental Table 1. The contigs and proteins used in this study are also available through figshare (10.6084/m9.figshare.6499781). Genomes from Tully *et al*. (2017, 2018) that were manually refined have been updated in NCBI with the corresponding accession IDs: NZKR02000000, NZKQ02000000, NZJY02000000, PAEM02000000, PADP02000000, PAUS02000000, PAMN02000000, PBGP02000000, PBGL02000000, NHGH02000000. All supplemental data is available through figshare (10.6084/m9.figshare.6499781).

Acknowledgements

I would like to acknowledge and thank Drs. Rohan Sachdeva, Johanna Holm, and Sarah Hu for reading, commenting, and enhancing drafts of this manuscript. Elaina Graham provided invaluable support for running various bioinformatic pipelines. A special thanks to Dr. John Heidelberg for the suggestion of a Hobbit-based naming schema. I would like to thank the Center for Dark Energy Biosphere Investigations (C-DEBI) for funding (OCE-0939654). And as I have noted before in previous research, I am grateful for the commitment of the *Tara* Oceans consortium to providing open access to their expansive metagenomic dataset.

References

1. DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 5685–5689 (1992).
2. Massana, R., Murray, A. E., Preston, C. M. & DeLong, E. F. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl. Environ. Microbiol.* **63,** 50–56 (1997).
3. Teira, E., Reinthaler, T., Pernthaler, A., Pernthaler, J. & Herndl, G. J. Combining catalyzed reporter deposition-fluorescence in situ hybridization and microautoradiography to detect substrate utilization by bacteria and Archaea in the deep ocean. *Appl. Environ. Microbiol.* **70,** 4411–4414 (2004).
4. Murray, A. E. *et al.* Time series assessment of planktonic archaeal variability in the Santa Barbara Channel. *Aquat. Microb. Ecol.* **20,** 129–145 (1999).
5. Pernthaler, A., Preston, C. M., Pernthaler, J., DeLong, E. F. & Amann, R. Comparison of fluorescently labeled oligonucleotide and polynucleotide probes for the detection of pelagic marine bacteria and archaea. *Appl. Environ. Microbiol.* **68,** 661–667 (2002).
6. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348,** –1262073 (2015).
7. Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology* **1,** 1–7 (2016).
8. Orsi, W. D. *et al.* Ecophysiology of uncultivated marine euryarchaea is linked to particulate organic matter. **9,** 1747–1763 (2015).
9. Martin-Cuadrado, A.-B. *et al.* A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. *ISME J* **9,** 1619–1634 (2015).
10. Hugoni, M. *et al.* Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences* **110,** 6004–6009 (2013).
11. Frigaard, N.-U., Martinez, A., Mincer, T. J. & DeLong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847–850 (2006).
12. Iverson, V. *et al.* Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* **335,** 587–590 (2012).
13. Baker, B. J. *et al.* Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J* **7,** 1962–1973 (2013).
14. Deschamps, P., Zivanovic, Y., Moreira, D., Rodriguez-Valera, F. & López-García, P. Pangenome Evidence for Extensive Interdomain Horizontal Transfer Affecting Lineage Core and Shell Genes in Uncultured Planktonic Thaumarchaeota and Euryarchaeota. *Genome Biology and Evolution* **6,** 1549–1563 (2014).
15. Li, M. *et al.* Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. *Nature Communications* **6,** 1–6 (2015).
16. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5,** e3558–15 (2017).
17. Tully, B. J., Graham, E. D., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5,** 170203 (2018).

18. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2,** 1–10 (2017).
19. Delmont T. O., Quince C., Shaiber A., Esen Ö. C., Lee S. T. M., Rappé M. S., MacLellan S. L., Lücker S., Eren A. M. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology* **326,** 1–12 (2018).
20. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4,** 170093–20 (2017).
21. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2,** 150023 (2015).
22. Thrash, J. C. *et al.* Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico 'Dead Zone'. *mBio* **8,** e01017–17–20 (2017).
23. Haro-Moreno, J. M. *et al.* Fine Stratification Of Microbial Communities Through A Metagenomic Profile Of The Photic Zone. *bioRxiv* 1–30 (2017). doi:10.1101/134635
24. Haroon, M. F. *et al.* A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci. Data* **3,** 160050 (2016).
25. Adam, P. S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. *Nature* **11,** 2407–2425 (2017).
26. Spang, A., Caceres, E. F. & Ettema, T. J. G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357,** (2017).
27. Galand, P. E., Gutiérrez-Provecho, C., Massana, R., Gasol, J. M. & Casamayor, E. O. Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnol. Oceanogr.* **55,** 2117–2125 (2010).
28. Grüber, G., Manimekalai, M. S. S., Mayer, F. & Müller, V. ATP synthases from archaea: The beauty of a molecular motor. *BBA - Bioenergetics* **1837,** 940–952 (2014).
29. Schlegel, K., Leone, V., Faraldo-Gómez, J. D. & Muller, V. Promiscuous archaeal ATP synthase concurrently coupled to Na+ and H+ translocation. *Proceedings of the National Academy of Sciences* **109,** 947–952 (2012).
30. Pinhassi, J., DeLong, E. F., Béjà, O., González, J. M. & Pedrós-Alió, C. **Marine Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology**. *Microbiol. Mol. Biol. Rev.* **80,** 929–954 (2016).
31. Localized high abundance of Marine Group II archaea in the subtropical Pearl River Estuary: implications for their niche adaptation. *Environ. Microbiol.* **20,** 734–754 (2017).
32. Borges, N. *et al.* Mannosylglycerate: structural analysis of biosynthesis and evolutionary history. *Extremophiles* **18,** 835–852 (2014).
33. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348,** 1261359–1261359 (2015).
34. Kopf, A. *et al.* The ocean sampling day consortium. *GigaScience* **4,** 27 (2015).
35. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28,** 15–18 (2000).
36. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,** D344–8 (2006).
37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
38. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 1–10 (2017). doi:10.1038/s41564-017-0012-7
39. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* **350,** 1–6 (2018).
40. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5,** e3035–19 (2017).

41. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Meth* **11,** 1144–1146 (2014).
42. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3,** e1319 (2015).
43. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
44. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1,** 16048 (2016).
45. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
46. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242–W245 (2016).
47. Bateman, A. *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **30,** 276–280 (2002).
48. Rawlings, N. D., Waller, M., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **42,** D503–D509 (2013).
49. Petersen, T. N., Brunak, S., Heijne, von, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth* **8,** 785–786 (2011).
50. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37,** D233–D238 (2009).
51. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40,** W445–W451 (2012).
52. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31,** 371–373 (2003).
53. Boeuf, D., Audic, S., Brillet-Guéguen, L., Caron, C. & Jeanthon, C. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015,** bav080–8 (2015).
54. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* **5,** e11147 (2010).
55. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421–9 (2009).
56. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **6,** e4320–23 (2018).
57. Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J. & Price, N. D. ITEP: an integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* **15,** 8 (2014).
58. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804,** 281–295 (2012).
59. Kopf, A. *et al.* The ocean sampling day consortium. *GigaScience* **4,** 27 (2015).
60. Li, H., Handsaker, B., Fennell, T., Ruan, J. & Homer, N. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
61. Hammer, Ø., Harper, D. & Ryan, P. D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4,** 9 (2001).

Supplemental Information

Supplemental Table 1. Information for all genomes used in study, including source ID numbers, family and genus assignment, completion stats (length, percent complete, percent contamination, percent strain heterogeneity), and genomes determined to be duplicates.

Supplemental Table 2. Counts of the number of proteins identified as an extracellular peptidase or carbohydrate-active enzyme for each genome.

Supplemental Table 3. A breakdown of the peptidases from the MEROPS database with corresponding IDs and Pfams.

Supplemental Table 4. All genomes with identified archaeal flagellum components, including the number of identified components and a prediction of if a full operon is present. Genomes from Tully *et al*. (2018) used to visualize the putative operon have NCBI contig accession and operon protein IDs listed.

Supplemental Table 5. Protein clusters generated using the Anvi'o pangenome workflow identified as either 'Core Thalassoarchaea', 'Core Delongarchiales', or 'Core Valerarchiales'. Only includes proteins with putative KEGG, CAZy, MEROPS, and Pfam assignments.

Supplemental Table 6. Corresponding metadata (environmental parameters) for *Tara* Oceans and Ocean Sampling Day samples.

Supplemental Data 1. FASTA format of the 16 ribosomal marker proteins for all *Thalassoarchaea* genomes.

Supplemental Data 2. Newick file of the phylogenomic tree generated using 16 concatenated ribosomal marker proteins for all *Thalassoarchaea* genomes. Corresponds to Supplemental Figure 1.

Supplemental Data 3. Newick file of the phylogenomic tree generated using 16 concatenated ribosomal marker proteins for the non-redundant set of *Thalassoarchaea* genomes. Corresponds to Figure 1.

Supplemental Data 4. Spreadsheet of the pairwise amino acid identity values.

Supplemental Data 5. FASTA format of the 16S rRNA gene sequences present in the *Thalassoarchaea* genomes.

Supplemental Data 6. Newick file of the phylogenetic tree generated using the 16S rRNA gene sequences present in the *Thalassoarchaea* genomes. Corresponds to Supplemental Figure 5.

Supplemental Data 7. Pfam HMMs used to identify MEROPS peptidases. The link between Pfams and MEROPS can be found in Supplemental Table 3.

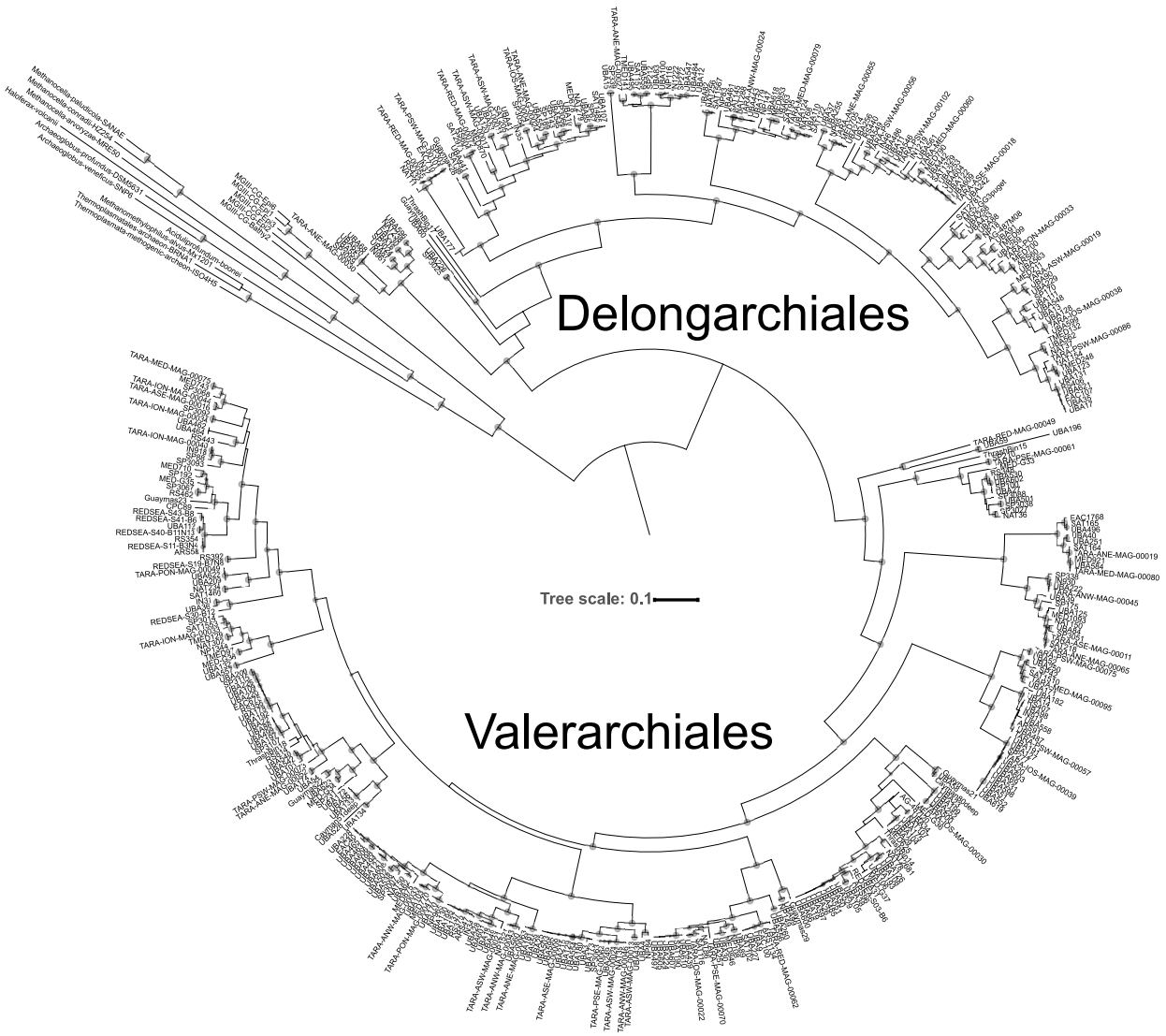
Supplemental Data 8. FASTA format of the putative proteorhodopsin sequences identified in the *Thalassoarchaea* genomes.

Supplemental Data 9. Newick file of the phylogenetic tree generated using the proteorhodopsin sequences present in the *Thalassoarchaea* genomes. Corresponds to Supplemental Figure 7.

Supplemental Data 10. GenBank files gathered from NCBI for the putative archaeal flagellum operon for Tully *et al*. (2018) genomes.

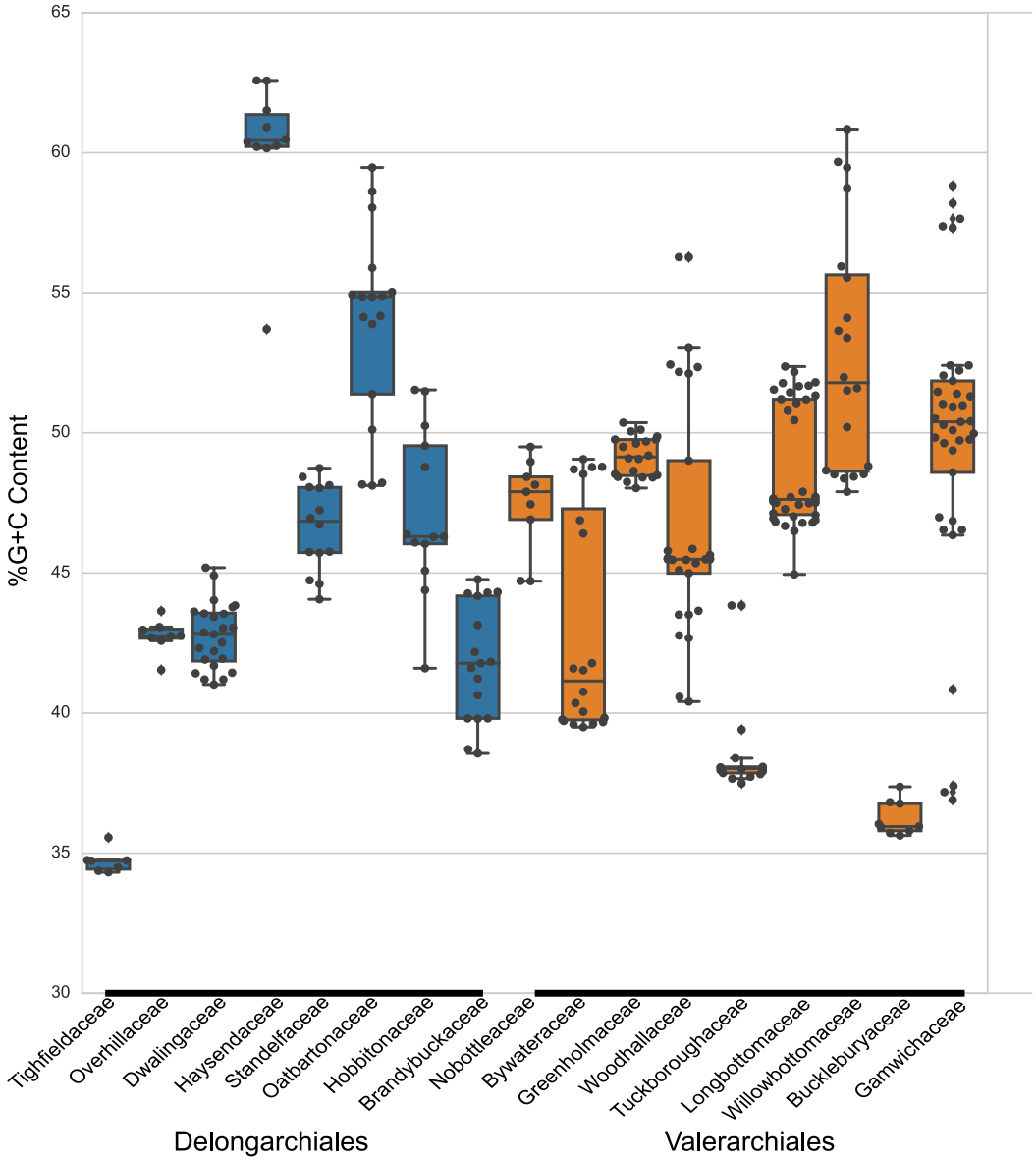
Supplemental Data 11. Raw read counts for all thalassoarchaeal contigs generated from *Tara* Oceans and Ocean Sampling Day samples.

Supplemental Data 12. Derived relative fraction and RPKM values for all *Thalassoarchaea* genomes for all *Tara* Oceans and Ocean Sampling Day samples. Corresponds to Figure 7 and Supplemental Figure 8.

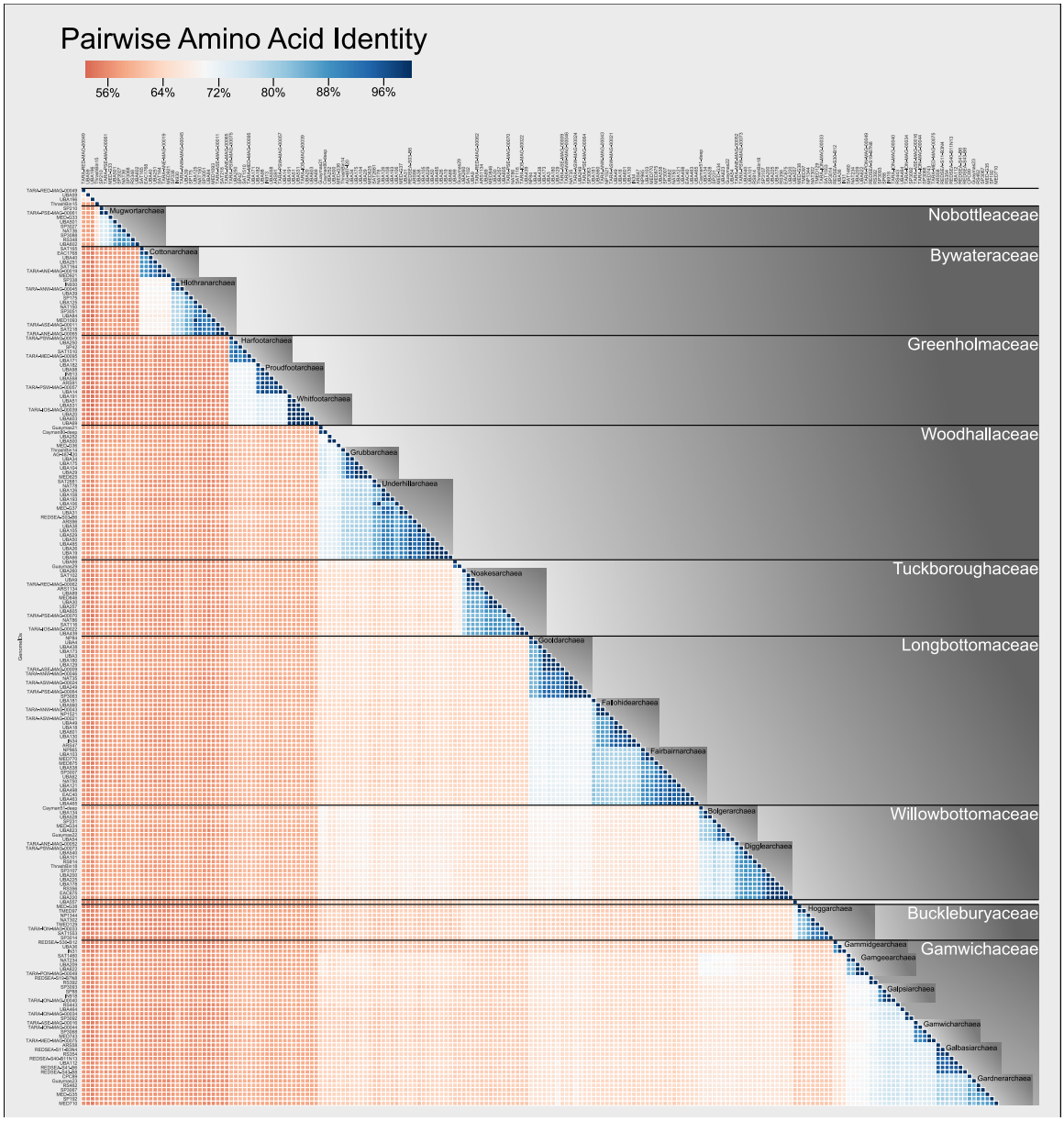


680

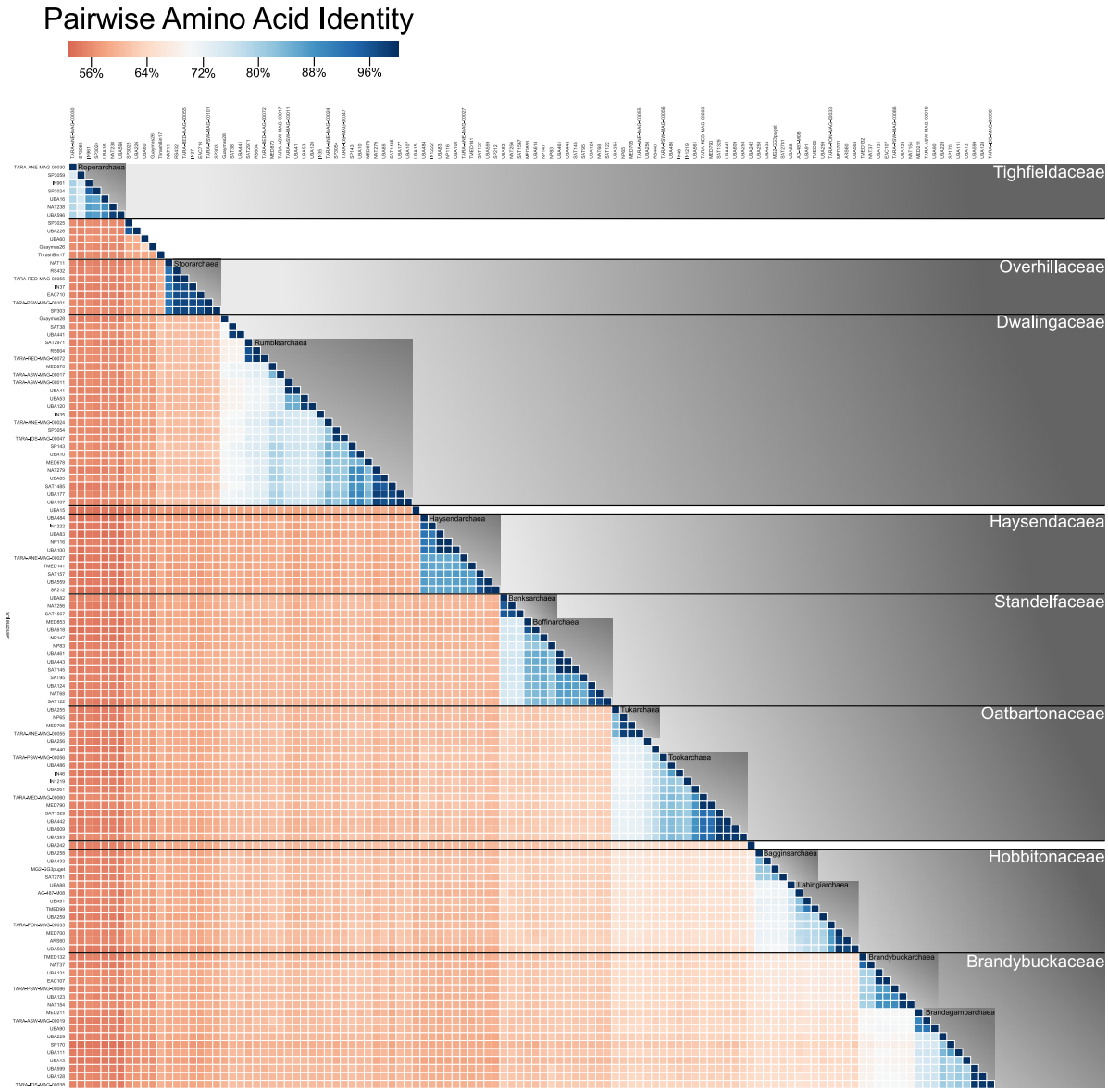
Supplemental Figure 1. A phylogenomic tree constructed using 16 concatenated ribosomal marker proteins for all of the *Thalassoarchaea*. Bootstrap values are scaled proportionally between 0.75-1.



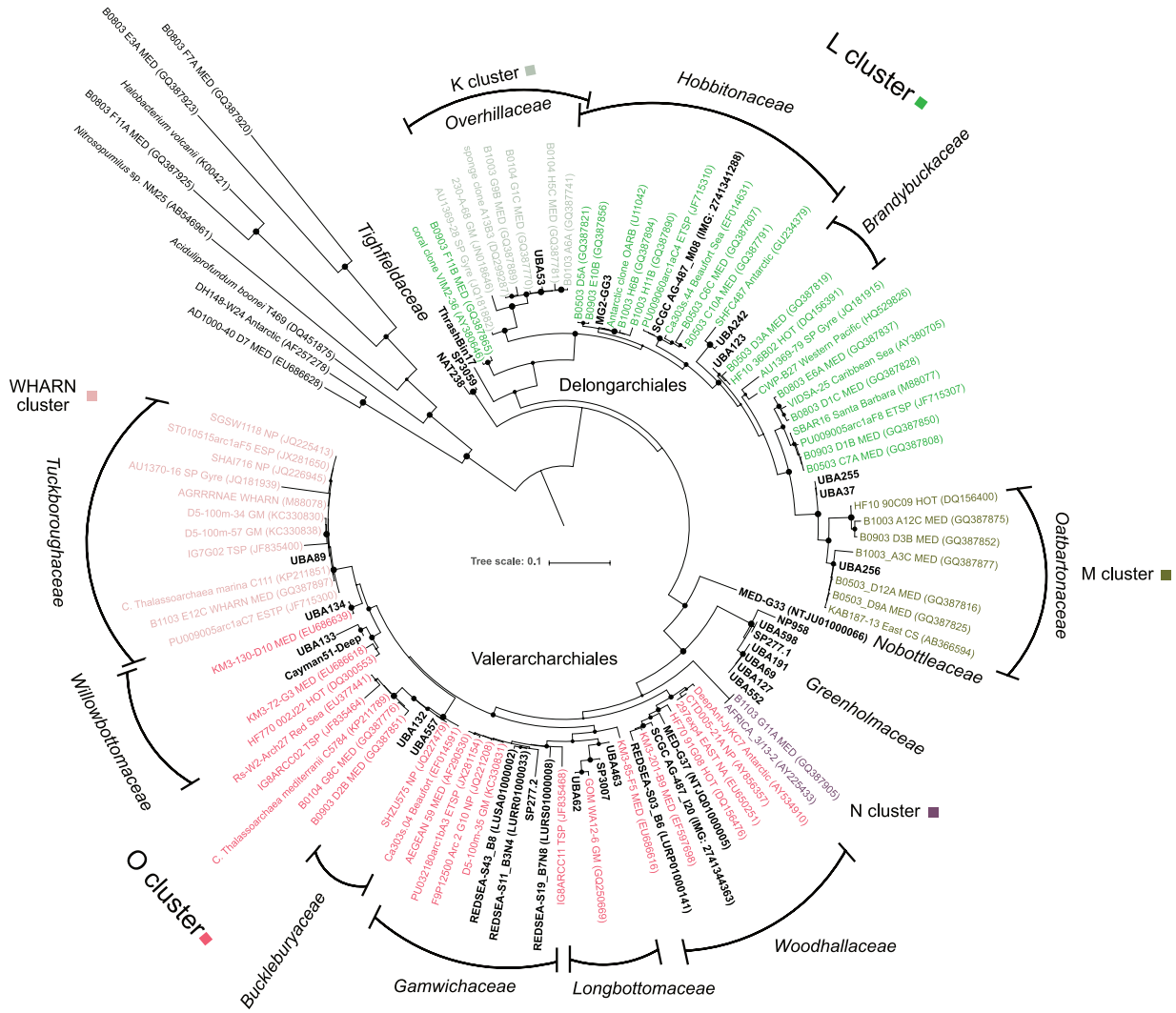
685 Supplemental Figure 2. Box plots illustrating the distribution of genome GC content at the family level.



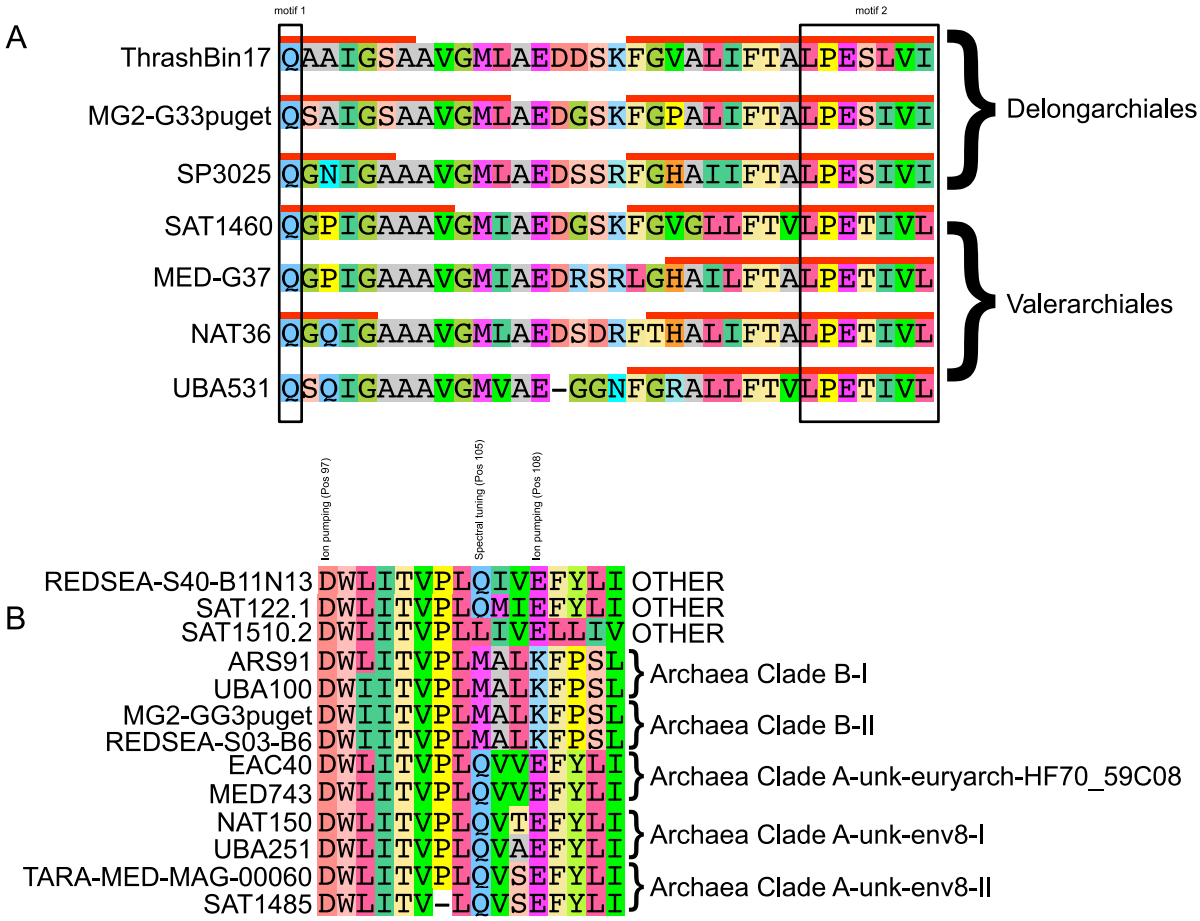
Supplemental Figure 3. Heatmap of pairwise amino acid identity for the *Delongarchiales*. Genomes are ordered based on their placement in the phylogenetic tree in Figure 1. Distinctions between families and genera are highlighted in gray. Genomes that could not be assigned to a family are highlighted in white.



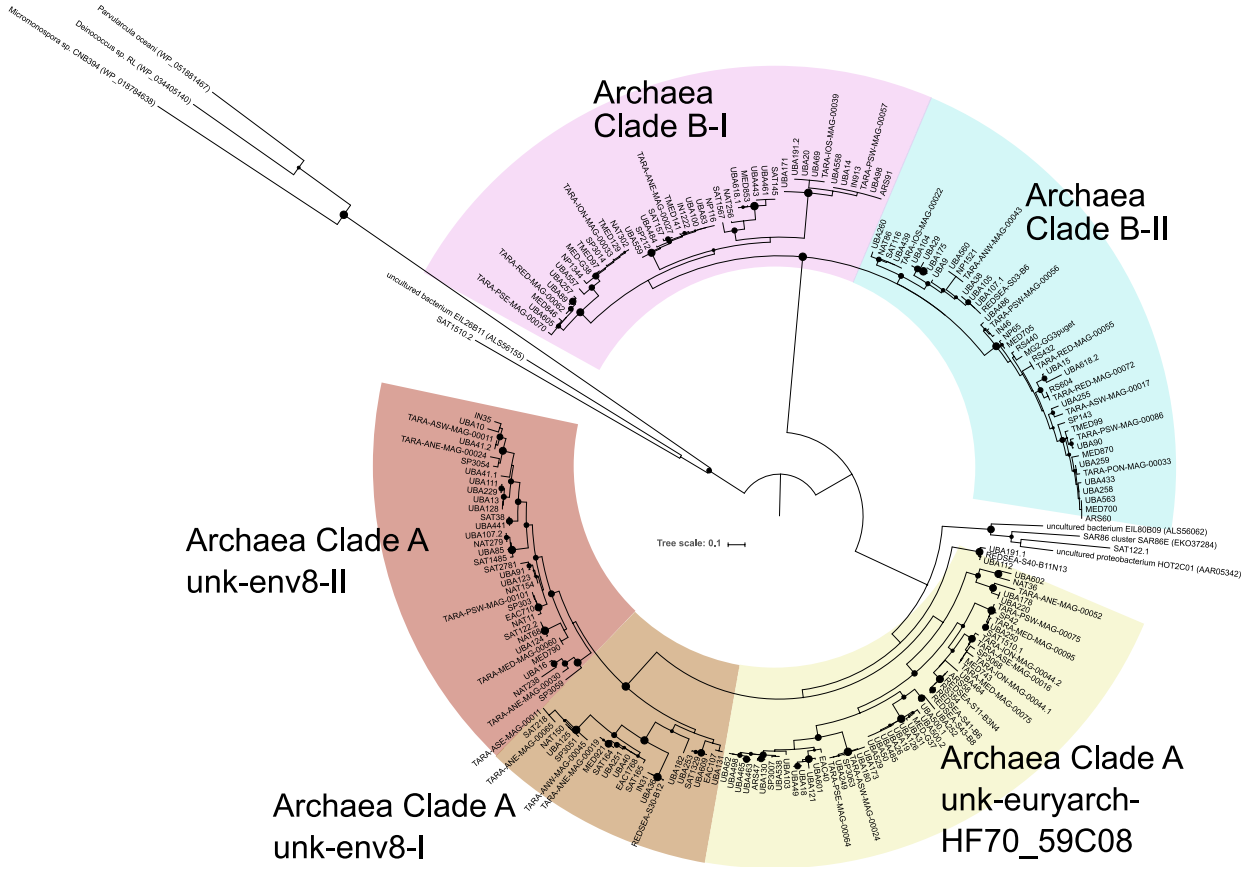
Supplemental Figure 4. Heatmap of pairwise amino acid identity for the *Valerarchiales*. Genomes are ordered based on their placement in the phylogenetic tree in Figure 1. Distinctions between families and genera are highlighted in gray. Genomes that could not be assigned to a family are highlighted in white.



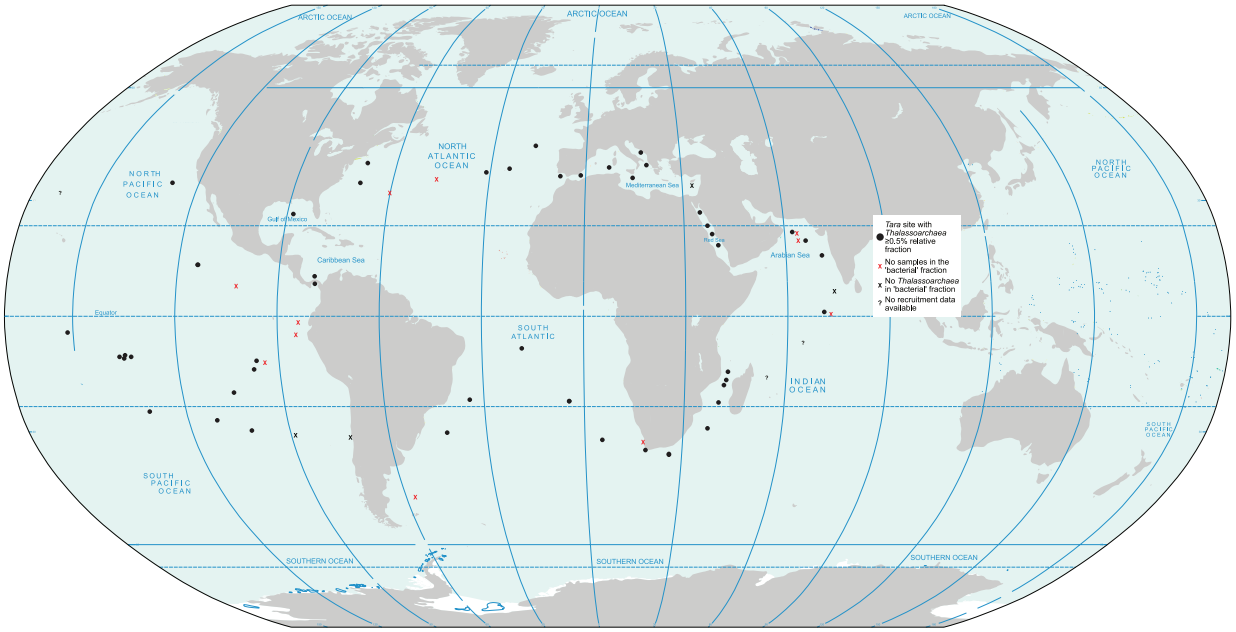
Supplemental Figure 5. A phylogenetic tree of the 16S rRNA gene for 35 thalassoarchaeal genomes combined with previously defined reference sequences. Previously observed clusters (denoted in various colors) that could be linked to newly defined families based on the occurrence of genome linked 16S rRNA sequences are shown. Internal nodes within large clusters (*e.g.*, L cluster) are used to demarcate different families where appropriate.



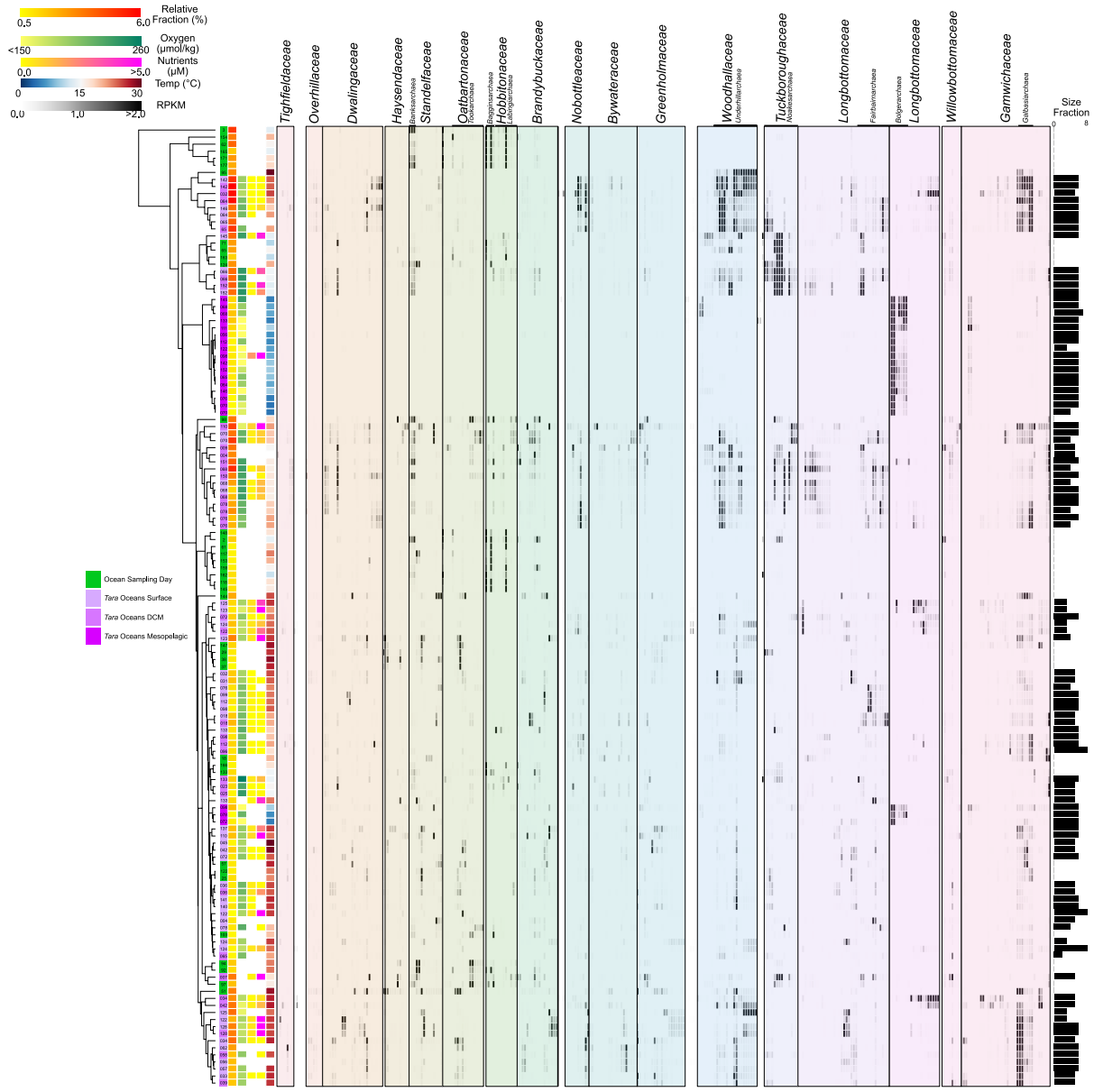
Supplemental Figure 6. A) Alignment *c* ring protein subunit (AtpK) for a selection of genomes. Transmembrane helices predicted using the TMHMM server (v.2.0) are denoted as red lines above the predicted region. Black boxes highlight the two conserved motifs that have been previously identified in Na^+ translocating ATP synthases. B) Alignment of the region used to predict functionality and spectral tuning amongst rhodopsins for a selection of genomes. Group assignments are based on clusters in Supplemental Figure 7. "OTHER" refers to rhodopsin sequences that did not fall within Archaea Clade A or B.



Supplemental Figure 7. Phylogenetic tree of the rhodopsin predicted proteins (190AA alignment) identified in the *Thalassoarchaea* genomes. Subclades in previously established clades are denoted. Bootstrap values are scaled proportionally between 0.75-1.



Supplemental Figure 8. Global map detailing the locations of the *Tara* Oceans sampling stations. Stations where at least one metagenomic sample recruited ≥0.5% relative fraction to the thalassoarchaeal genomes are represented as black dots. Stations lacking metagenomic samples in the 'bacterial' size fraction (0.22-3.0µm) are denoted as a red 'X'. Stations with metagenomic samples in the 'bacterial' size fraction but did not recruit ≥0.5% relative fraction are denoted as a black 'X'. Three stations (TARA048, -052, and -132) were not included in this analysis are denoted as question marks.



Supplemental Figure 9. A heatmap displaying the RPKM values for all of the *Thalassoarchaea* genomes discussed in the manuscript in high abundance samples (≥0.5% relative fraction). RPKM values are scaled from 0-2 with values ≥2 in black (median, 0.001; maximum, 31.54). Samples are hierarchically clustered based on all *Thalassoarchaea* RPKM values and the sample source is displayed as either green (OSD) or purple (*Tara* Oceans). Numbers displayed for samples correspond to sample/station ID. The available environmental parameters are presented as colored heatmaps (missing parameters are displayed as white). The size fraction graph has a range of 0-8: 0, whole water sample ≥0.22µm (OSD samples only); 1, <0.22µm; 2, 0.1-0.22µm; 3, 0.1-0.8µm; 4, 0.22-0.8µm; 5, 0.22-1.6µm; 6, 0.22-3µm; 7, 0.45-0.8µm; 8, 0.8-5.0µm. The order of the sample hierarchical clustering and displayed environmental parameters are the same as those presented in Figure 5.