Metabolic Diversity within the Globally Abundant Marine Group II *Euryarchaea* Offers Insight into 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 clades, the MGII have distinct ecological patterns that are not understood based on the limited number of available genomes. Here, I present the comparative genomic analysis of 250 MGII genomes, providing the most detailed view of these mesophilic archaea to-date. This analysis identified 17 distinct subclades including nine subclades that previously lacked reference genomes. The metabolic potential and distribution of the MGII genera revealed distinct roles in the environment, identifying algal-saccharide-degrading coastal subclades, protein-degrading oligotrophic surface ocean subclades, and mesopelagic subclades lacking proteorhodopsins common in all other subclades. 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 *Euryarchaea* (MGII) have remained an enigmatic group of organisms in the marine environment. The MGII have been identified at high abundance in surface oceans[2,3] and can account for ~15% of the archaeal cells in the oligotrophic open ocean[4]. The MGII have been shown to increase in abundance in response to phytoplankton blooms[5] and can comprise up to ~30% of the total microbial community after a bloom terminates[6]. Research has shown that the MGII correlate with specific genera of phytoplankton[7], during and after blooms[8], can be associated with particles when samples are size fractionated[9], and correlate with a novel clade of marine viruses[10]. Phylogenetic analyses have revealed the presence of two dominant clades of MGII, referred to as MGIIa and MGIIb (recently *Thalassoarchaea* has been proposed as a name for the MGIIb[11]), that respond to different environmental conditions, including temperature and nutrients[12].

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 is derived from interpretations of environmental sampling data (*e.g.*, 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[13], 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[11,14-16], genes representative of particle-attachment[9,14], and genes for the biosynthesis of tetraether lipids[11,17]. Comparatively, the capacity for motility via archaeal flagellum has only been identified in some of the recovered genomes[11,14]. Much of this primary literature is reviewed in Ref. 18.

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[19-22] provide an

avenue for exploring the metabolic variation between the MGIIa and MGIIb, and in conjunction with environmental data collected from the same filter fractions and sampling depths[23,24] can be used to understand the variables and conditions that favor each clade. Here, the analysis of 250 MGII genomes identifies the metabolic traits unique to the genomes derived from the MGIIa and MGIIb, 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 subclades, 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. 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[11,14,17,25,26]. 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[17,25,26] (16 genomes) with only six genomes for the MGIIa available[14,17,25]. A combined 409 genomes reconstructed from marine environmental metagenomes, originating from four studies utilizing the *Tara* Oceans dataset (designations TMED[19], TOBG[20], UBA[21], and TARA-MAG[22]) and the Red Sea (designated as REDSEA[27]), were identified in publicly available databases. As the metagenome-assembled genomes generated from studies using the *Tara* Oceans data were generated from the same metagenomes, it was important to identify an potentially identical genomes within the recovered genome collection. Groups of genomes were identified that shared ≥98.5% average nucleotide identity (ANI) and a representative was selected for each group based on completion and contamination metrics. In the past, ~96% ANI has been used to denote organisms of different species[28]. A cutoff of 98.5% ANI was selected as to only exclude identical genomes, while attempting to retain genomes with strain level differences sampled from different metagenomes. A total of 90 groups of at least two genomes were identified (Supplemental Data 1) and removal of redundant genomes resulted in a dataset of 258 non-redundant genomes. A phylogenetic tree using 120 concatenated single copy marker proteins was constructed using all genomes that had ≥60 single copy markers (Figure 1). Eight genomes had an insufficient number of markers and were no longer considered for further analysis (a tree with the same parameters was applied to the redundant genome dataset and included 410 genomes; Supplemental Figure 1; Supplemental Table 1). The 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 subclades – 8 MGIIa subclades (designated MGIIa.1-8) and 9 MGIIb subclades (designated MGIIb.9-17). The 17 subclades were further supported by pairwise ANI and average amino acid identity (AAI; Supplemental Figure 2 and 3; Supplemental Data 2). Nine of the clades were composed exclusively of genomes reconstructed from the *Tara* Oceans metagenomic dataset. Based on the extrapolated genome size for these 17 subclades, MGIIa genome sizes were significantly larger than MGIIb genomes, on average ~400kbp (Figure 2A; $p << 0.001$). 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 percent GC (%G+C) content and phylogenetic group. %G+C content of the genomes had a wide range of values (~35%->60%; Figure 2B). Additionally, several subclades had high internal variation of %G+C content.

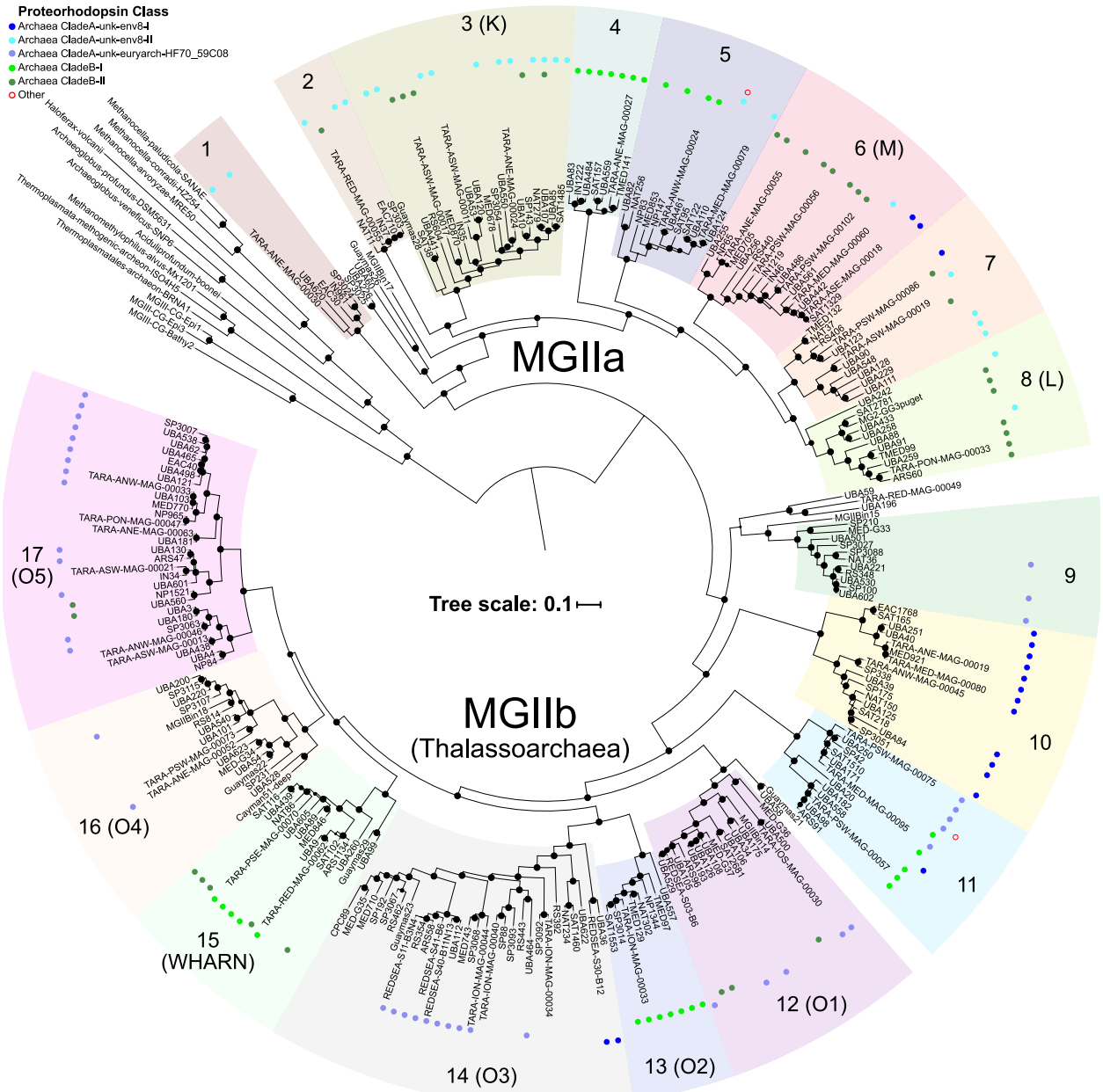


Figure 1 A phylogenomic tree constructed using 120 concatenated single copy marker proteins for the MGII. Subclades are denoted based on relative distance from the root. Parentheses indicate subclades that can be attributed to previously defined 16S rRNA gene clusters. Bootstrap values are scaled proportionally 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 MGII genomes had 16S rRNA gene sequences (n = 34) which were used to determine the relationship between previously identified sequence clusters[11,29] and the newly identified subclades (Supplemental Figure 4). 16S rRNA sequences from the MGIIa.1 and MGIIb.9 were not represented in previously identified MGII 16S rRNA gene clusters. Conversely, the previously identified N cluster did not have representative amongst the environmental genomes, either as a result of missing diversity among the described genomes or due to the fact that not all of the subclades had a representative with an identified 16S rRNA gene. Several of defined 16S rRNA gene clusters could be linked directly to subclades with genomic representatives based on phylogeny; the WHARN cluster branched with a representative of the MGIIb.15 and the M and L clusters branched with the MGIIa.6 and MGIIa.8,

respectively. The O cluster from the MGIIb, could be divided at several internal nodes and assigned to five of the subclades. The phylogeny of several of the representative 16S rRNA sequences associated with the O cluster could not be fully resolved likely as the result of a low quality sequence across the length of the alignment or due to an incorrect placement within the genome during binning.

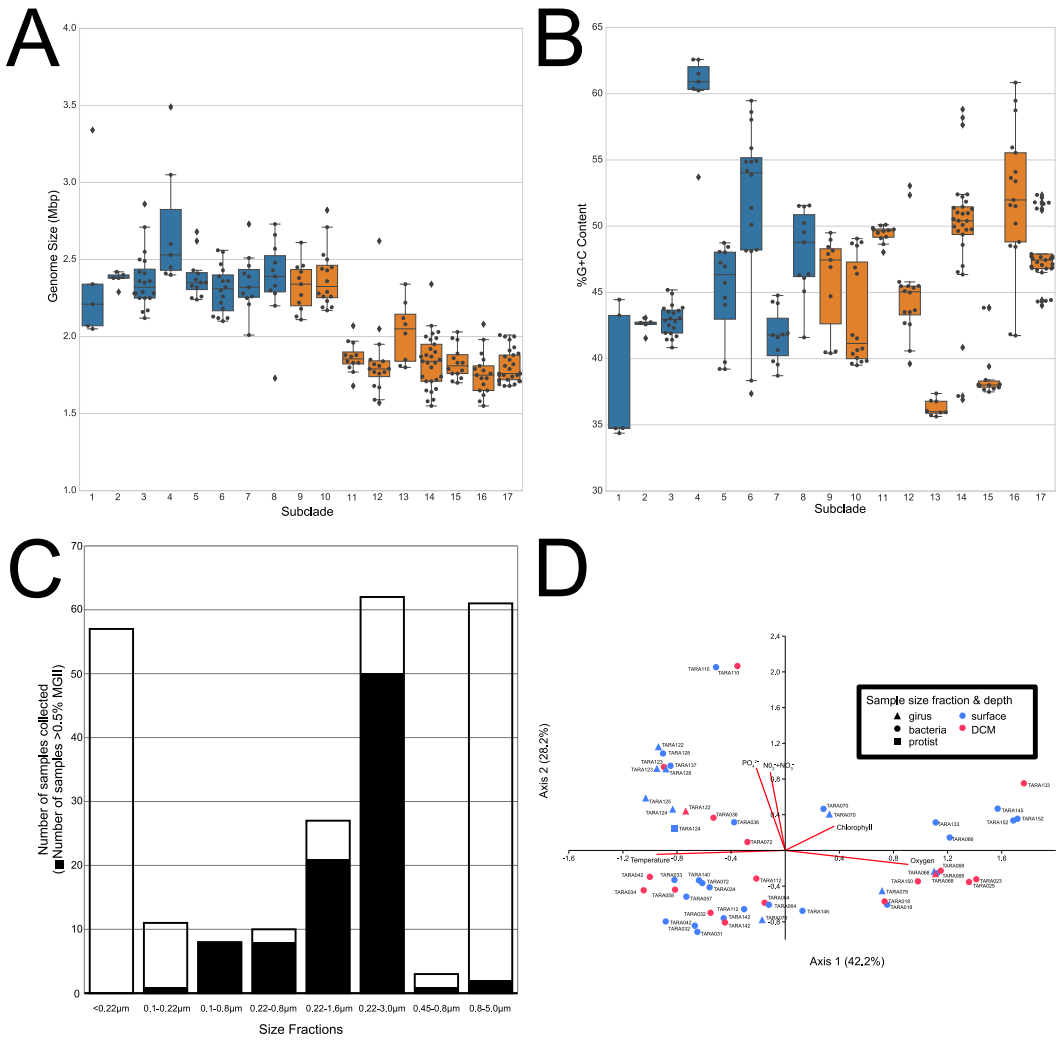


Figure 2 *A. Box plots illustrating the distribution of genome sizes for each subclade (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; diamonds, outliers). MGIIa and MGIIb box plots are displayed in blue and orange, respectively. B. Box plots illustrating the distribution of genome %G+C content for each subclade. MGIIa and MGIIb box plots are displayed in blue and orange, respectively. 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 MGII. D. A canonical correspondence analysis (CCA) of the high abundance (≥0.5% relative fraction) samples (n = 54) based on the RPKM values for the MGII genomes.*

MGII share an electron transport chain with putative Na^+ pumping components. There were several shared traits amongst the MGIIa and MGIIb, particularly related to the components of the electron transport chain (ETC). Genomes belonging to both groups had canonical NADH dehydrogenases (complex I) and succinate dehydrogenases (complex II) that link electron transport to oxygen as a terminal electron acceptor via low-affinity cytochrome *c* oxidases (Figure 3). As has been noted previously[9], most members of the MGII possessed genes encoding a cytochrome *b* and a Rieske iron-sulfur domain protein but lacked the genes for the canonical cytochrome bc_1 (complex III). Many of the MGII 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 MGII lacked the subunits needed to complete the complex (RnfACDEG). Thus, it was surprising that distributed across all of the subclades in 188 genomes (75%), the MGII 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[30] (Supplemental Figure 5A). The motif in the second helix appears to be diagnostic of the clade a genome belongs to: the MGIIa contained a LPESxxI motif and the MGIIb contained a LPETIxL motif. The presence of these motifs does not preclude ATP synthesis via H^+ pumping[31], though a majority of the experimentally confirmed A_1A_O ATP synthases with these motifs exclusively pump Na^+ ions[30].

MGII share the ability to degrade extracellular proteins and fatty acids. As has been reported previously[11,14-16], a majority of the MGII 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 MGII 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 MGIIa or MGIIb are also part of the MGII core genome[11,17], 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).

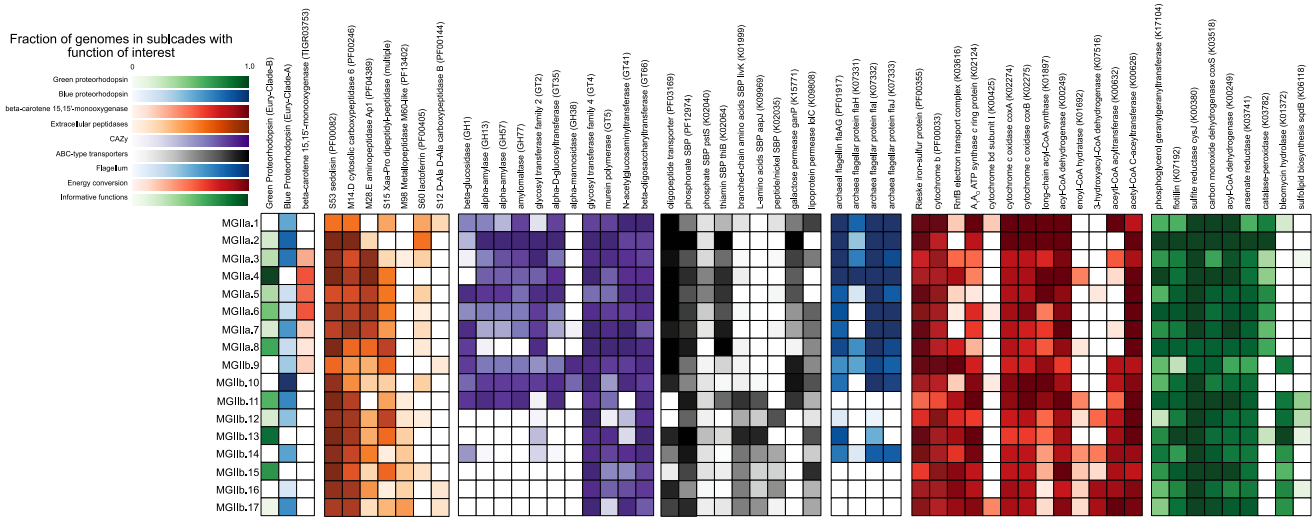


Figure 3. Heatmap of the occurrence of various functions of interest in the respective MGII subclades. Heatmaps are scaled from 0 to 1, where 1 represents that all genomes within the designated subclade possess the function of interest.

Putative proteorhodopsins differentiate members of the MGIIa and MGIIb. Components of the ETC and HMW degradation were present in all the subclades, there were several traits that either lacked a phylogenetic signature or differentiated the MGIIa and the MGIIb. As has been noted previously[14] and confirmed with this collection of genomes, taxa in all of the MGII subclades 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 (PRs)[32] capable of establishing H^+ gradients (Supplemental Figure 5B). Phylogenetically, these PRs cluster in

established clades[33] Archaea Clade A (Clade-A) and Archaea Clade B (Clade-B) and based on the amino acid in position 105[34] (glutamine/methionine), spectral tuning[35] prediction indicates sensitivity to blue and green light, respectively (Supplemental Figure 5B). Five subclades exclusively possess Clade-A, three subclades exclusively possess Clade-B, and nine subclades have genomes that possess either of the two PRs.

A group of genomes from MGIIb.16, including a number of genomes reconstructed or present (see below) in the aphotic zone, do not possess PRs (Figure 1). The lack of PRs in deep-sea MGII is consistent across the tree, with deep-sea reconstructed genomes not within the MGIIb.16 tend to represent the most basal branching members of other subclades (*e.g.*, genome Guaymas21 basal to MGIIb.12). PRs 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 as identified in the MICrhoDE database, while Clade-B has two distinct groups (Clade-B-I and -II) (Supplemental Figure 6). The MGIIa possessed all of the PR groups, except unk-euryarch-HF70_59C08 and slightly favor the green light tuned PRs (55% of PR containing genomes), while the MGIIb do not utilize the CladeA-unk-env8-II group and favor blue light tuned PRs (67% of PR containing genomes). Additionally, several subclades 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 MGII 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 subclades from the MGIIa and all but one of the subclades from the MGIIb lacked this crucial functional step.

The degradation of extracellular peptidases and algal oligosaccharides differentiate members of the MGIIa and MGIIb. While the MGII shared several functionalities with a role in the degradation of HMW organic matter, there was a greater diversity of functionality in specific clades and subclades (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 19 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 can be processed by an organism. Three of the five extracellular peptidase classes were distributed across both the MGIIa and MGIIb, while the M60-like metallopeptidase and carboxypeptidase B, were present almost exclusively amongst the MGIIb. Despite sharing many of the putative protein degrading functions, subclades from the MGIIb, except for MGIIb.9 and MGIIb.10, 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 MGIIa only have the previously noted oligopeptide transporter (Figure 3).

Beyond the degradation of proteins and fatty acids, there is evidence to suggest that MGII have a role in the degradation of carbohydrate HMW organic matter[36]. Interestingly, glycoside hydrolases with functionality for the degradation of algal oligosaccharides, including pectin, starch, and glycogen, are found exclusively amongst the MGIIa and the most basal families of the MGIIb, the MGIIb.9, MGIIb.10, and MGIIb.11 (Figure 3). These same subclades also possess an annotated galactose permease subunit for an ABC-type transporter. Further, MGIIb.9 and MGIIb.10 also possess a glycoside hydrolase that could possibly play a role in mannosylglycerate degradation, an osmolyte found in red algae[37].

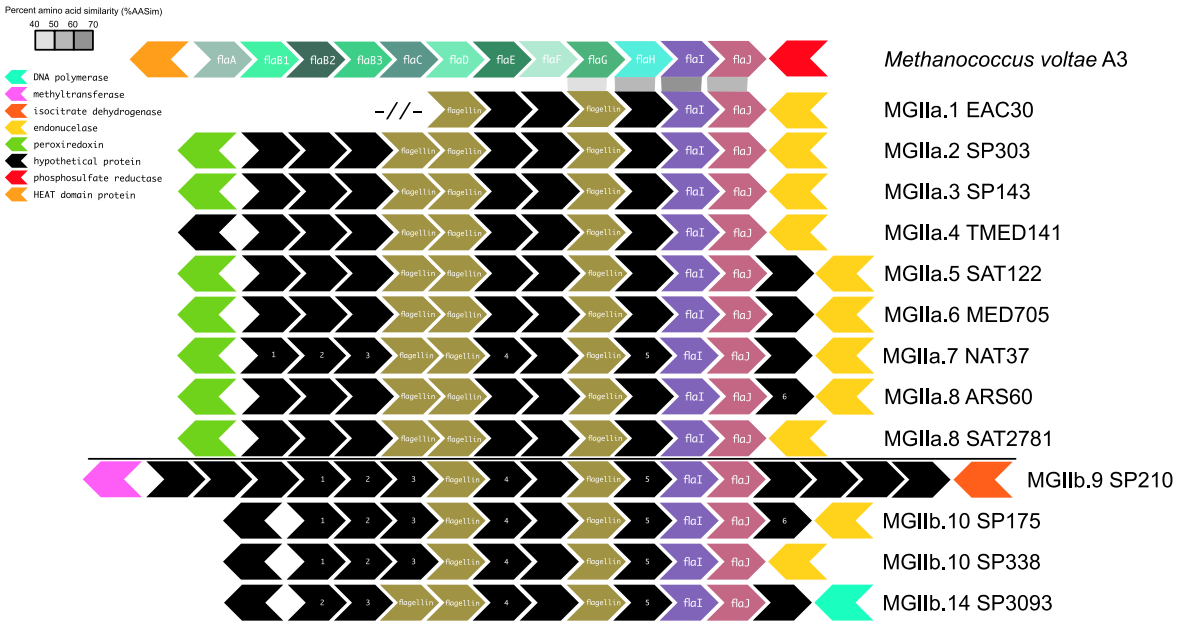


Figure 4. A stylized view of the putative motility genes present in the MGII. The longest contig for each subclade is shown. Hypothetical proteins lacking KEGG annotations or matches to queried HMMs are in black. All hypothetical proteins in a column had significant BLAST matches to their neighbors, except as noted by the numbers in the transition from the MGIIa to the MGIIb. 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 MGIIa.1 genome are noted.

Motility is a trait common to the MGIIa. Previous research has shown evidence for and against the putative capacity for motility amongst the MGII[11,14]. The analyzed genomes lacked annotations or homology for most of the canonical archaeal flagellum operon (Figure 4; Supplemental Table 3). However, genomes from all of the MGIIa subclades, MGIIb.9, MGIIb.10, and MGIIb.14 possessed proteins annotated as subunits from the canonical operon (FlaAGHIJ). A comparison of the identified subunits from a representative of the MGIIa.1 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 a similarity comparison 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 MGIIa, while some variation in gene content could be observed amongst the subclades from the MGIIb.

For several other functions ascribed to the MGII as a whole[11], there are distinct distributions amongst the clades, including the presence of a catalase-peroxidase amongst the MGIIa and a bleomycin hydrolase amongst the MGIIb (Figure 3). Further, several other predicted metabolic functions appear to be specific to only a subset of subclades and may have a role in niche differentiation, 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).

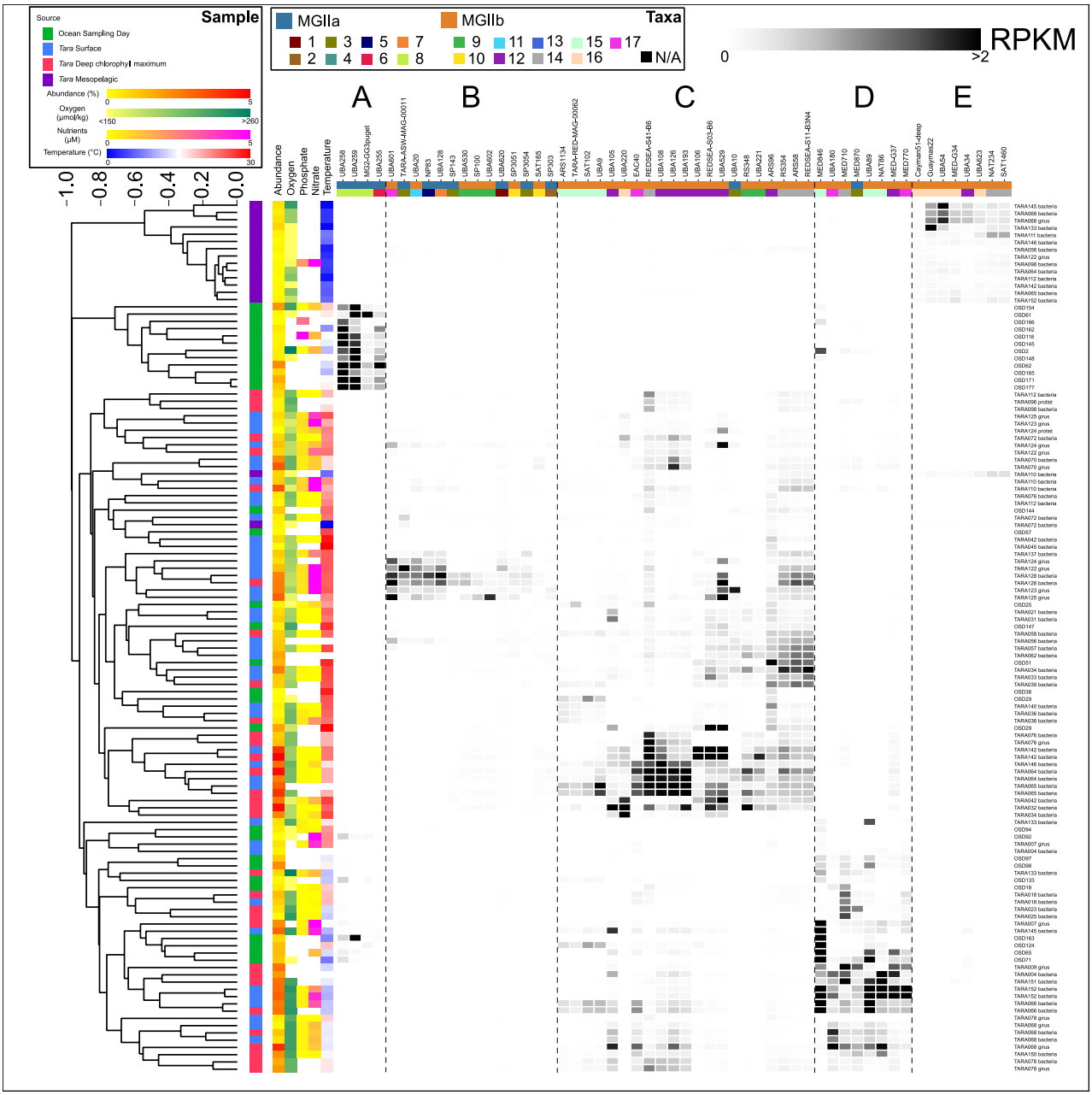


Figure 5. A heatmap displaying the RPKM values for a subset of MGII genomes discussed in the manuscript in high abundance samples (≥0.5% relative fraction). RPKM values are scaled from 0-2 with values ≥2 in black (median, 0.0012; maximum, 13.09). Samples are hierarchically clustered based on all MGII RPKM values and the sample source is displayed. The available environmental parameters are presented as colored heatmaps (missing parameters are not displayed). The heatmap used to identify high abundance clusters is available in Supplemental Figure 8.

MGII subclades belong to ecological clusters with unique distributions. Using a comprehensive set of *Tara* Oceans metagenomic datasets from across the globe[24], that included all of the size fractions for which DNA was collected (viral, 'bacterial', and eukaryotic), it was possible to explore where specific MGII groups were dominant. The MGII were rarely found to be abundant (>0.5% relative abundance; mean, 1.75%; maximum, 4.97%) 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 MGII were abundant at all *Tara* Oceans stations with a 'bacterial' size fraction (n = 45), except for six stations

(Supplemental Figure 7). There were no *Tara* Oceans metagenomic samples collected from size fractions >5µm. A canonical correspondence analysis (CCA) based on the MGII community structure for 54 samples divided samples along a gradient of temperature and oxygen concentrations (x-axis, 42.2% observed variance) and on nutrient concentrations (y-axis, 28.2% observed variance; Figure 2D). Overall, samples from the surface and deep chlorophyll maximum (DCM) were intermingled in all quadrants of the CCA plot suggesting that incidence light availability may not be a contributing factor to community composition. Interestingly, while the MGII are not persistently abundant in the smallest *Tara* Oceans size fractions, of the 11 girus size fraction samples on the CCA plot, a majority (55%) group together correlated to high temperature and phosphate and nitrate concentrations.

The abundance of the MGIIa in open ocean samples was limited. In an effort to identify samples where the clade may be abundant, 118 'prokaryotic' metagenomes from coastal (<10km) Ocean Sampling Day[38] 2014 (OSD) samples were assessed for the presence of the MGII 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 MGII in the 'bacterial' *Tara* Oceans fractions, only about a quarter of the samples (n = 31) from OSD had high MGII abundance.

Tara Oceans and OSD samples and taxa were hierarchically clustered based on community composition using Bray-Curtis distances. Five groups of taxa that occurred at high abundance in a discreet set of samples were identified (Supplemental Figure 8). These groups were classified as "Ecological clusters" A-E, an indication of the shared ecology between the taxa. Ecological cluster A consisted of a group of four MGIIa, predominantly from MGIIa.8, that were abundant exclusively in samples from the OSD (Figure 5). Though coastal, these OSD samples were marine in nature (30-38 practical salinity unit [PSU] compared to *Tara* Oceans samples 29-40 PSU). The lack of consistency in the types of data collected from OSD samples limits the statistical power for interpreting the influence of oxygen and nutrients concentration on these taxa ($p_{BH\text{-}FDR} < 0.02$), but these OSD samples have moderate temperatures (12-21°C) compared to many of the other samples ($F_{PERMANOVA} = 12.04$, $p_{BH\text{-}FDR} = 0.0001$). Ecological cluster E consisted of a group of taxa that were exclusively abundant in mesopelagic samples. Though the cluster does exclude UBA528, which is the most abundant taxa in all mesopelagic samples (Supplemental Figure 8). This cluster is entirely composed of taxa from the MGIIb, with a majority from the MGIIb.16. All but one mesopelagic sample lack nutrient data, but Ecological cluster E was significantly correlated to low oxygen concentrations (3-237 µmol/kg, $F_{PERMANOVA} = 2.884$, $p_{BH\text{-}FDR} = 0.0002$) and low temperatures (<10.8°C, $F_{PERMANOVA} = 7.95$, $p_{BH\text{-}FDR} = 0.0002$).

Ecological clusters B-D were less confined to a single sample type, but correlated with specific environmental conditions. Ecological cluster B consisted of 14 taxa from both the MGIIa (n = 7) and MGIIb (n = 7) that were abundant predominantly in surface samples with warm temperatures (24-27°C, $F_{PERMANOVA} = 9.559$, $p_{BH\text{-}FDR} = 0.0002$), moderate oxygen (160-195 µmol/kg, $F_{PERMANOVA} = 4.315$, $p_{BH\text{-}FDR} = 0.0002$), low phosphate (0.5-0.6 µM, $F_{PERMANOVA} = 2.872$, $p_{BH\text{-}FDR} = 0.0003$), and high nitrate (2.4-5.5 µM, $F_{PERMANOVA} = 2.003$, $p_{BH\text{-}FDR} = 0.0015$) concentrations. Ecological cluster C consists of 21 taxa almost exclusively from the MGIIb ($n_{MGIIb} = 20$) with high abundance in samples from both the surface and DCM, though members of cluster are distributed among all of the high temperature samples (19-31°C, $F_{PERMANOVA} = 7.579$, $p_{BH\text{-}FDR} = 0.0002$). Many of these samples also possessed moderate oxygen (171-212 µmol/kg, $F_{PERMANOVA} = 3.564$, $p_{BH\text{-}FDR} = 0.0002$), low phosphate (<0.5 µM, $F_{PERMANOVA} = 1.918$, $p_{BH\text{-}FDR} = 0.003$), and low nitrate (<2.1 µM, $F_{PERMANOVA} = 1.38$, $p_{BH\text{-}FDR} = 0.0491$) concentrations. Ecological cluster D was predominantly MGIIb taxa ($n_{MGIIb} = 7$) from samples with moderate temperatures (11-21°C, $F_{PERMANOVA} = 11.36$, $p_{BH\text{-}FDR} = 0.0002$), high oxygen (>231 µmol/kg, $F_{PERMANOVA} = 4.99$, $p_{BH\text{-}FDR} = 0.0002$), and low phosphate (<0.4 µM, $F_{PERMANOVA} = 1.714$, $p_{BH\text{-}FDR} = 0.0212$) concentrations. Ecological cluster D was not statistically correlated with nitrate concentrations.

Discussion

The details provided by the increased resolution of MGII genomes collected for this study in phylogeny, metabolism, and ecology redefines what is understood about these globally dominant

mesophilic *Euryarchaea*. Previous phylogenetic diversity contained within reconstructed genomes and genomic fragments failed to capture at least nine newly identified subclades. This collection of 250 genomes allows for a more precise understanding of the metabolic potential present in the MGII, including the metabolic and ecological differentiation of the MGIIa and MGIIb.

There are several unifying elements in the overall predicted MGII metabolism. All MGII have the metabolic potential to be obligate aerobic heterotrophs with many possessing the capacity to harass solar energy via PRs. There is no collective indication that the MGII have the capacity for anaerobic metabolisms, though members of four subclades (MGIIa.1, MGIIa.6, MGIIa.10, and MGIIb.17) have the genomic potential to function in suboxic conditions through the presence of a high-affinity cytochrome *bd* oxidase. Central to the potential heterotrophic metabolism is the degradation of HMW proteins and fatty acids that can exploited by a number of transporters and extracellular peptidases. Energy conversion via oxidative phosphorylation uses a number of canonical (complex I and II) and MGII specific (complex III) electron transfer proteins, but the MGII may also be capable of utilizing a sodium motive force through an A_1A_O ATP synthase. A Na^+-driven ATP synthases differentiates the MGII from most marine bacteria that possess F_1F_O H^+-driven ATP synthases. How this MGII H^+/Na^+ 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 MGII utilize both H^+ and Na^+, similar to *Methanosarcinales* under marine conditions[31], and that different elements of the MGII ETC perform these translocations.

The presence of PRs is ubiquitous across the MGII, with members of the MGIIa and MGIIb predicted to have spectral tuning to both blue (Clade-A) and green (Clade-B) light. But while neither PR clade adheres to a strict MGII phylogenetic distribution, there are specific patterns that may provide insight in to the adaptation of certain subclades or taxa to localized conditions. For example, MGIIa.4, MGIIb.13, and MGIIb.15 possess only green spectral tuned PRs. While taxa from MGIIa.4 and MGIIb.13 individually do not reach high relative abundance, abundant taxa from MGIIb.15 are present in Ecological clusters C and D and are correlated to open ocean samples with distinct temperature and oxygen profiles. The taxa abundant in the two Ecological clusters are phylogenetically distinct within the MGIIb.15 subclade, suggesting fine-scale genomic adaptation may play a role how these organisms are distributed in the ocean. Further, MGII genomes reconstructed from the bathypelagic (Guaymas-derived genomes), abyssopelagic (Cayman-derived genome), and identified as abundant in mesopelagic *Tara* Oceans samples (Ecological cluster E) all lack PRs indicating that MGII heterotrophy is not dependent on a light-amended proton motive force.

There are two important metabolic distinctions between the MGIIa and most of the MGIIb – the be potential for the degradation of oligosaccharides and flagellum-based motility. Reported in the analysis of the first MGII genome[14], the potential for oligosaccharide degradation has often been ascribed to all members of the MGII. Analysis of the presence of glycoside hydrolases in MGII genomes suggests that this metabolism is limited to the MGIIa and the three most basal subclades of the MGIIb. Further, the specificity of the annotated glycoside hydrolases suggests that the MGII are degrading algal derived substrates. With the exception of MGIIb.11, the subclades with the capacity to degrade algal oligosaccharides also have the potential for motility via a proposed archaeal flagellum-based system. The capacity for motility may be required to exploit ephemeral sources of oligosaccharides through chemotaxis between algal cells/detritus. This metabolic link to algal oligosaccharides may limit the abundance of these subclades to coastal environments, such as MGIIa subclades in Ecological cluster A, or during/post phytoplankton blooms, as previously observed[39].

Many of the individual taxa identified in this collection of 250 genomes are not abundant in any one sample, but the aggregated MGII community dynamics revealed the importance of temperature and oxygen availability to shape the individual constituents of a sample. Both temperature and oxygen availability directly impact the heterotrophic metabolisms of the MGII through the requirement of oxygen by aerobes for carbon degradation and the role temperature plays in enhancing reaction kinetics at warmer temperatures. Combined with measurements for phosphate, nitrate, and chlorophyll, these components of the environment explained more than two-thirds (70%) of the community variance. While chlorophyll

measurements capture one of the important biotic interactions, it is likely that the missing variance can/will be explained by MGII-microbe and MGII-virus interactions.

The predicted metabolic functions observed in the subclades appear to correlate with observations of MGII taxa and community constituents in global samples. The most abundant MGIIa genomes occur in coastal samples originating from OSD samples and not the open ocean *Tara* metagenomes. These coastal conditions would correlate to areas of the oceans where algal oligosaccharides would be more readily available. Additionally, high nitrate concentrations can stimulate growth of phytoplankton, which may explain why Ecological cluster B has a high MGIIa abundance from open ocean samples collected in the Equatorial Pacific Ocean (TARA122-125 and TARA128). Conversely, the MGIIb dominant ecological clusters that occur in environments where direct algal inputs are limited (*i.e.*, open ocean and mesopelagic samples). There are several ecological clusters for which the current metabolic predictions cannot explain. Difficulty in this regard is partially a result of the limited degree of functional annotation achieved for the MGII (~45% KEGG functional annotation efficiency). Ecological clusters C and D are composed predominantly of MGIIb subclades and have unique distributions based on non-overlapping temperature and oxygen concentrations. The subclades that compose each Ecological cluster are similar, suggesting that intra-subclade genomic variations may explain this distinct ecology.

Concluding Remarks

There remain several important avenues of research that cannot be addressed despite the scale of the genome dataset. The MGII have previously been identified in filter fractions greater than 3µm and were hypothesized to have been attached to large plankton[9]. Virtually none of the *Tara* Oceans samples in the 3-5µm range recruited to the MGII genomes, but the lack of metagenomes from >5µm from *Tara* Oceans makes this difficult to assess if the MGII may be attached to even larger particles. And while global metagenomes provide an metric for asking if an organism present, it fails to capture what an organism doing. The release of the corresponding *Tara* Oceans microbial metatranscriptomes will allow future assessment of what metabolisms are active for the MGII. With the identification of a putatively Na^+-driven ATP synthase, future research will be required to determine how the MGII ETC functions, especially with the prevalence of H^+-pumping proteorhodopsins. Should the MGII ETC obligately rely on Na^+ motive force, this may explain 'limited' distribution of MGII in marine environments, unlike the more global distribution of the *Thaumarchaeota* (formerly Marine Group I). The analysis of MGII genomic potential across 17 subclades allows for the reinterpretation of the role these organisms play in the cycling of HMW organic matter in the environment and opens these new avenues of research.

Methods

Genome Selection and Phylogenetic Assessment

MGII genomes that were publicly available prior to January 1, 2018[14,17,25-27] were collected from NCBI[40] and IMG[41] and were assessed using CheckM[42] (v1.0.11) to determine the approximate completeness and degree of a contamination (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[19,20,22,43] 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*.[44] (2018). Briefly, high contamination genomes originally binned using BinSanity[45] (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[46] (v.0.4.1; -c 800 -I 500). Putative bins determined by CONCOCT were identified that contained contigs from the BinSanity genomes. Overlapping bins and read coverage data from associated samples were profiled (anvi-profile), combined (anvi-merge) and visualized in Anvi'o[47] (v.3) (anvi-interactive). The putative MGII genomes were manually refined based on the hierarchically clustered coverage profiles and %G+C. The Anvi'o profiles used to perform this refinement for the analyzed genomes (see below) are available (https://doi.org/10.6084/m9.figshare.7154813.v1) and further details pertaining to the refinement method can be found in the Supplemental Information.

402 A pairwise comparison of percent average nucleotide identity (ANI) was performed on the initial collection
403 of redundant, putative MGII genomes (n = 431) to determine which genomes that may represent identical genomic
404 data (fastANI[28]; --fragLen 1500). Groups of genomes with ≥98.5% ANI were determined to be identical and the
405 genome with the highest completion and lowest contamination estimates were selected as a nonredundant
406 representative for further analysis (Supplemental Data 1; n = 258). Putative coding DNA sequence (CDS) were
407 predicted for each using Prodigal[35] (v.2.6.3). The predicted proteins sequences for each genome were searched
408 (HMMER[48] v.3.1b2; hmmsearch -E 1E-5) using HMM models representing the 120 single copy marker proteins as
409 detailed by the Genome Taxonomy Database (GTDB v.86.0; http://gtdb.ecogenomic.org/) and sourced from
410 TIGRfam[49] (v.15) and Pfam[50] (v.31.0) (Supplemental Data 3). All proteins with a match to a single copy marker
411 model were aligned using MUSCLE[51] (v.3.8.31; -maxiters 8) and automatically trimmed using trimAL[39]
412 (v.1.2rev59; -automated1). Genomes with ≥60 single copy marker proteins were retained for further analysis (n =
413 250; Supplemental Data 3). Proteins alignments were concatenated and a phylogenomic tree was constructed using
414 FastTreeDbl[40] (v.2.1.10; -gamma -lg; Supplemental Data 4 and 5). All described phylogenetic trees were visualized
415 using the Interactive Tree of Life[52] (v.4.2.3). Pairwise amino acid identity (AAI) was calculated for the genomes,
416 separated by into the major MGII clades (MGIIa and MGIIb) using CompareM
417 (https://github.com/dparks1134/CompareM; v.0.0.23; aai_wf defaults; Supplemental Figure 2 and 3; Supplemental
418 Data 2). Subclades 1-17 were defined based on a cutoff of 1.2 relative phylogenetic distance from the root and
419 supported by clustering of pairwise relationships >70% ANI and >~70% AAI.
420 All genomes, including those identified as redundant[19-22,27], were assessed for the presence of the 16S
421 rRNA gene using RNAmmer[42] (v.1.2; -S arch -m ssu; Supplemental Data 6). Identified sequences were combined
422 with 16S rRNA gene sequences representing the available various reference genomes[14,17,25,26] and previously
423 established clusters[11] (MGIIA clusters K, L, M; MGIIB clusters O, N, WHARN). As above, sequences were aligned
424 using MUSCLE, and then manually trimmed to remove gaps in Geneious (v.6.1.8). A 430 bp alignment was used to
425 construct a phylogenetic tree using FastTree (-nt -gtr -gamma; Supplemental Figure 4; Supplemental Data 7). When
426 possible, the previously defined 16S rRNA gene clusters were related back to the subclades identified on the single
427 copy marker protein tree.
428
429 **Functional Prediction**
430 A uniform functional annotation was applied to all predicted proteins for the set of analyzed genomes.
431 Proteins were annotated with the KEGG database[43] using GhostKOALA[44] ('genus_prokaryotes+family_eukaryotes';
432 accessed September 17, 2018; v.2.0). Extracellular peptidases (enzymes predicted to degrade proteins) were
433 identified with matches (hmmsearch -T 75) to PFAM HMM models[53] corresponding to MEROPS peptidase
434 families[54] (Supplemental Table 4; Supplemental Data 8) that were predicted to have "extracellular" or "outer
435 membrane" localization by PSortb[47] (v.3; -a) or an "unknown" localization with predicted translocation signal
436 peptides by SignalP[55] (v.4.1; -t gram+). Carbohydrate-active enzymes (CAZy)[56] were identified (hmmsearch -T 75)
437 using HMM models from dbCAN[57] (v.6). Functions of interest were predominantly identified based on the
438 corresponding KEGG Orthology (KO) entry and GhostKOALA predictions. Specific functions of interest without a
439 KO entry were searched using HMM models (hmmsearch -T 75) obtained from PFAM and TIGRFAM.
440 Predicted proteins of each genome were screened for matches to the rhodopsin PFAM model (PF01036;
441 hmmsearch -T 75; Supplemental Data 9). In order to identify putative proteorhodopsins, sequences matching the
442 rhodopsin HMM model were processed using the Galaxy-MICrhoDE workflow implemented on the Galaxy web
443 server (http://usegalaxy.org) to assign rhodopsins to the MICrhoDE database[58]. The alignment generated from the
444 workflow was manually trimmed to a 96 amino acid region conserved across all sequences, re-aligned using
445 MUSCLE and used to construct a phylogenetic tree with FastTree (as above; Supplemental Data 10). The
446 rhodopsins were predominantly assigned to three clades based on the phylogenetic relationships with other
447 MICrhoDE sequences, unk-euryarch-HF70-59C08, unk-env8, and one unassigned clade. Two rhodopsins were
448 assigned to additional clades, MICrhoDE clade IV-Proteo3-HF10_19P19 and another unassigned clade. Based on
449 Pinhassi *et al.* (2016), unk-euryarch-HF70-59C08 and unk-env8 are also known as Archaea Clade-A and the
450 unassigned clade belongs to Archaea Clade-B. A more detailed phylogenetic tree was constructed (as above) using
451 only sequences from MGII and a reduced outgroup (Supplemental Figure 6). The MGII rhodopsin sequences were
452 aligned using MUSCLE and were assessed for specific amino acids present at positions 97 and 108 to determine
453 putative function and position 105 to determine putative spectral tuning (Supplemental Figure 5B).
454 The operon putatively encoding an archaeal flagellum was identified based on the presence of co-localized
455 the flagellar proteins FlaHIJ (K07331-3) and archaeal flagellins (PF01917). All genomes with possible co-
456 localization of these proteins were identified (Supplemental Table 3). Putative operons from the analyzed TOBG
457 genomes were visualized by subclade using the progressiveMauve aligner[59] (v.2.3.1; default) and longest contig

458 containing the operon was selected to represent that subclade (Supplemental Data 11). Each representative was the
459 compared to its phylogenetic neighbor using BLASTP[60] (v.2.2.30+; parameters) to identify orthologs.
460
461 **MGII Core Genome Analysis**
462 A pangenomic analysis was performed for the analyzed genomes using the Anvi'o pangenome workflow[61]
463 (v.5). FASTA files for each genome were converted to a contigs database (anvi-script-FASTA-to-contigs-db) and
464 each was amended with their corresponding GhostKOALA KEGG annotations (anvi-import-functions). The
465 pangenomic analysis (anvi-pan-genome) used the default parameters for minbit[62] (--minbit 0.5) and MCL[63] (--mcl-
466 inflation 2) to generate protein clusters (PCs). Additional information regarding which clade and subclade each
467 genome belonged was amended to the pangenome database (anvi-import-misc-data). Calculations to determine
468 which functions were putatively enriched was performed at the clade (MGIIa or MGIIb) and subclade (1-17) levels
469 (anvi-get-enriched-functions-per-pan-group). Enriched functions were determined to be a putative core function for
470 the MGII if the corresponding protein cluster(s) was detected in ≥70% of all MGIIa and MGIIb genomes. Functions
471 core to the MGIIa or MGIIb were protein cluster(s) present in ≥70% of genomes within the clade and in ≤30% of the
472 genomes in the other clade. A function was determined to be putatively unique to a subclade if the protein cluster(s)
473 was present in ≥70% of genomes within the subclade and in ≤10% of all other genomes (Supplemental Table 5).
474
475 **MGII Relative Fraction and Environmental Correlations**
476 The analyzed set of MGII genomes were used to recruit sequences from environmental metagenomic
477 libraries, specifically 238 samples from *Tara* Oceans representing 62 stations and 118 samples from Ocean
478 Sampling Day (OSD) 2014[64] (Supplemental Table 6). Metagenomic sequences were recruited using Bowtie2[58]
479 (v.2.2.5; --no-unal). Resulting SAM files were sorted and converted to BAM files using SAMtools[65] (v.1.5; view;
480 sort). BAM files were filtered using BamM (https://github.com/minillinim/BamM; v.1.7.3) to select for reads with
481 ≥95% sequence identity with ≥75% alignment (--percentage_id 0.95 --percentage_aln 0.75). featureCounts[60]
482 (v.1.5.0-p2; default parameters) implemented through Binsanity-profile[45] (v.0.2.6.4; default parameters) was used to
483 generate read counts for each contig from the filtered BAM files (Supplemental Data 12). Read counts were used to
484 calculate the relative fraction of each analyzed MGII genome in all metagenomic samples (reads recruited to a
485 genome ÷ total reads in metagenomic sample) and reads per kbp of each genome per Mbp of each metagenomic
486 sample (RPKM; (reads recruited to a genome ÷ (length of genome in bp ÷ 1000)) ÷ (total bp in metagenome ÷
487 1000000)) (Supplemental Data 13). Samples were divided into high (≥0.5% MGII recruitment) and low relative
488 fraction samples (<0.5% MGII recruitment). Based on these designations, RPKM values for MGII genomes from
489 *Tara* Oceans samples with high relative fraction with sufficient metadata (depth, temperature, and oxygen,
490 chlorophyll, phosphate, and nitrate [measured as nitrate + nitrite]), were used in a canonical correspondence
491 analysis[66] (CCA) in Past3[67] (v.3.20). Due the correlation of depth with a number of factors, temperature,
492 chlorophyll, phosphate, and nitrate, depth measurements and one mesopelagic sample with all five parameters were
493 removed from the final CCA (data not shown). OSD samples consistently only collected temperature, distance from
494 the coast, and salinity and were not included in CCA analysis. RPKM values for MGII genomes from high relative
495 fraction samples were hierarchically clustered with average linkage clustering based on a Bray-Curtis distance
496 matrix for both samples and taxon separately. Clustering was implemented with SciPy (http://www.scipy.org;
497 v.1.0.0) and visualized with seaborn (http://seaborn.pydata.org; v.0.8.1). Ecological clusters of genomes were
498 identified based on clades within each dendrogram, defined as <0.8 dissimilarity in the taxon dendrogram and <0.7
499 dissimilarity in the sample dendrogram, that corresponded to elevated RPKM values (Supplemental Figure 8). These
500 ecological clusters were used to construct a reduced figure that maintained the sample and taxon clustering order
501 (Figure 5).
502
503 **Statistical Analysis**
504 Average genome length and %G+C for the MIIa and MGIIb were compared using a two-sample, two-tail
505 Student's t-test considering unequal variance.
506 Ecological clusters were correlated to temperature and oxygen, phosphate, and nitrate concentrations using
507 a one-way PERMANOVA[68] implemented in Past3. PERMANOVA F-statistics were determined based on 9,999
508 permutations performed on a Bray-Curtis[69] distance matrix. *p*-values were adjusted for multiple comparisons of
509 related dependent variables using the Benjamini-Hochberg False Discovery Rate[70] correction. Past3 implements an
510 *ad hoc* test of the pairwise PERMANOVA for all pairs of groups and corresponding *p*-values were adjusted with a
511 Bonferroni correction (Supplemental Data 14). The PERMANOVA groups for the dependent variables were:
512 temperature (3 groups), <10°C, 10-20°C, >20°C; oxygen (4 groups), <110, 110-159, 160-200, >200 μmol/kg;

phosphate (4 groups), below detection limit (BDL), <0.5, 0.5-2.0, >2.0 μM; nitrate (5 groups), BDL, <0.5, 0.5-1.9, 2.0-5.0, >5.0 μM.

Data Availability

The genomes used in this study are publicly available, except for a subset of deep-sea MGII from Guaymas basin presented in 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.7154813.v1). 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.7154813.v1).

Acknowledgements

I would like to acknowledge and thank Drs. Rohan Sachdeva and Sarah Hu for reading, commenting, and enhancing early 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 the (now defunct) Hobbit-based naming schema and acting as a combative reviewer. I would like to extend a special thank you to Dr. Johanna Holm for her patience when describing statistics and for proofreading early drafts of the manuscript. 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 and other researchers who have made their genomes so easily accessible.

References

1. DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 5685–5689 (1992).
2. Massana, R., DeLong, E. F. & Pedrós-Alió, C. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Appl. Environ. Microbiol.* **66,** 1777–1787 (2000).
3. 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).
4. 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).
5. Murray, A. E. *et al.* Time series assessment of planktonic archaeal variability in the Santa Barbara Channel. *Aquat. Microb. Ecol.* **20,** 129–145 (1999).
6. 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).
7. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348,** –1262073 (2015).
8. Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology* **1,** 1–7 (2016).
9. Orsi, W. D. *et al.* Ecophysiology of uncultivated marine euryarchaea is linked to particulate organic matter. **9,** 1747–1763 (2015).
10. Philosof, A. *et al.* Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota. *Current Biology* 1–8 (2017). doi:10.1016/j.cub.2017.03.052

11. 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).
12. 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).
13. 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).
14. Iverson, V. *et al.* Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* **335,** 587–590 (2012).
15. 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).
16. 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).
17. 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).
18. Zhang, C. L., Xie, W., Martin-Cuadrado, A.-B. & Rodriguez-Valera, F. Marine Group II Archaea, potentially important players in the global ocean carbon cycle. *Front. Microbio.* **6,** 2444–9 (2015).
19. 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).
20. 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).
21. 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).
22. Delmont T. O., *et al*. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology* **326,** 1–12 (2018).
23. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4,** 170093–20 (2017).
24. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2,** 150023 (2015).
25. 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).
26. 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
27. Haroon, M. F. *et al.* A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci. Data* **3,** 160050 (2016).
28. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High-throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *bioRxiv* 1–14 (2017). doi:10.1101/225342
29. 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).
30. 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).
31. 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).
32. Béjá, O. *et al.* Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289,** 1902–1906 (2000).

614 33. Pinhassi, J., DeLong, E. F., Béjà, O., González, J. M. & Pedrós-Alió, C. Marine Bacterial and
615 Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology. *Microbiol. Mol.*
616 *Biol. Rev.* **80,** 929–954 (2016).
617 34. Man, D. *et al.* Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* **22,** 1725–
618 1731 (2003).
619 35. Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin
620 phototrophy in the ocean. *Nature* **411,** 786–789 (2001).
621 36. Xie, W. *et al*. Localized high abundance of Marine Group II archaea in the subtropical Pearl River
622 Estuary: implications for their niche adaptation. *Environ. Microbiol.* **20,** 734–754 (2017).
623 37. Borges, N. *et al.* Mannosylglycerate: structural analysis of biosynthesis and evolutionary history.
624 *Extremophiles* **18,** 835–852 (2014).
625 38. Kopf, A. *et al.* The ocean sampling day consortium. *GigaScience* **4,** 27 (2015).
626 39. Needham, D. M. *et al.* Dynamics and interactions of highly resolved marine plankton via
627 automated high-frequency sampling. *ISME J* **347,** 1–16 (2018).
628 40. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28,** 15–18 (2000).
629 41. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,**
630 D344–8 (2006).
631 42. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing
632 the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome*
633 *Res.* **25,** 1043–1055 (2015).
634 43. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially
635 expands the tree of life. *Nature Microbiology* 1–10 (2017). doi:10.1038/s41564-017-0012-7
636 44. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-
637 distributed bacterial phototroph. *ISME J* **350,** 1–6 (2018).
638 45. Graham, E. D., Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering
639 of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5,** e3035–
640 19 (2017).
641 46. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Meth* **11,**
642 1144–1146 (2014).
643 47. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*
644 **3,** e1319 (2015).
645 48. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity
646 searching. *Nucleic Acids Res.* **39,** W29–W37 (2011).
647 49. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic*
648 *Acids Res.* **31,** 371–373 (2003).
649 50. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic*
650 *Acids Res.* **44,** D279–D285 (2016).
651 51. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
652 *Nucleic Acids Res.* **32,** 1792–1797 (2004).
653 52. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and
654 other trees. *Nucleic Acids Res.* **44,** W242–W245 (2016).
655 53. Bateman, A. *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **30,** 276–280 (2002).
656 54. Rawlings, N. D., Waller, M., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic
657 enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **42,** D503–D509 (2013).
658 55. Petersen, T. N., Brunak, S., Heijne, von, G. & Nielsen, H. SignalP 4.0: discriminating signal
659 peptides from transmembrane regions. *Nat Meth* **8,** 785–786 (2011).
660 56. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for
661 Glycogenomics. *Nucleic Acids Res.* **37,** D233–D238 (2009).
662 57. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation.
663 *Nucleic Acids Res.* **40,** W445–W451 (2012).

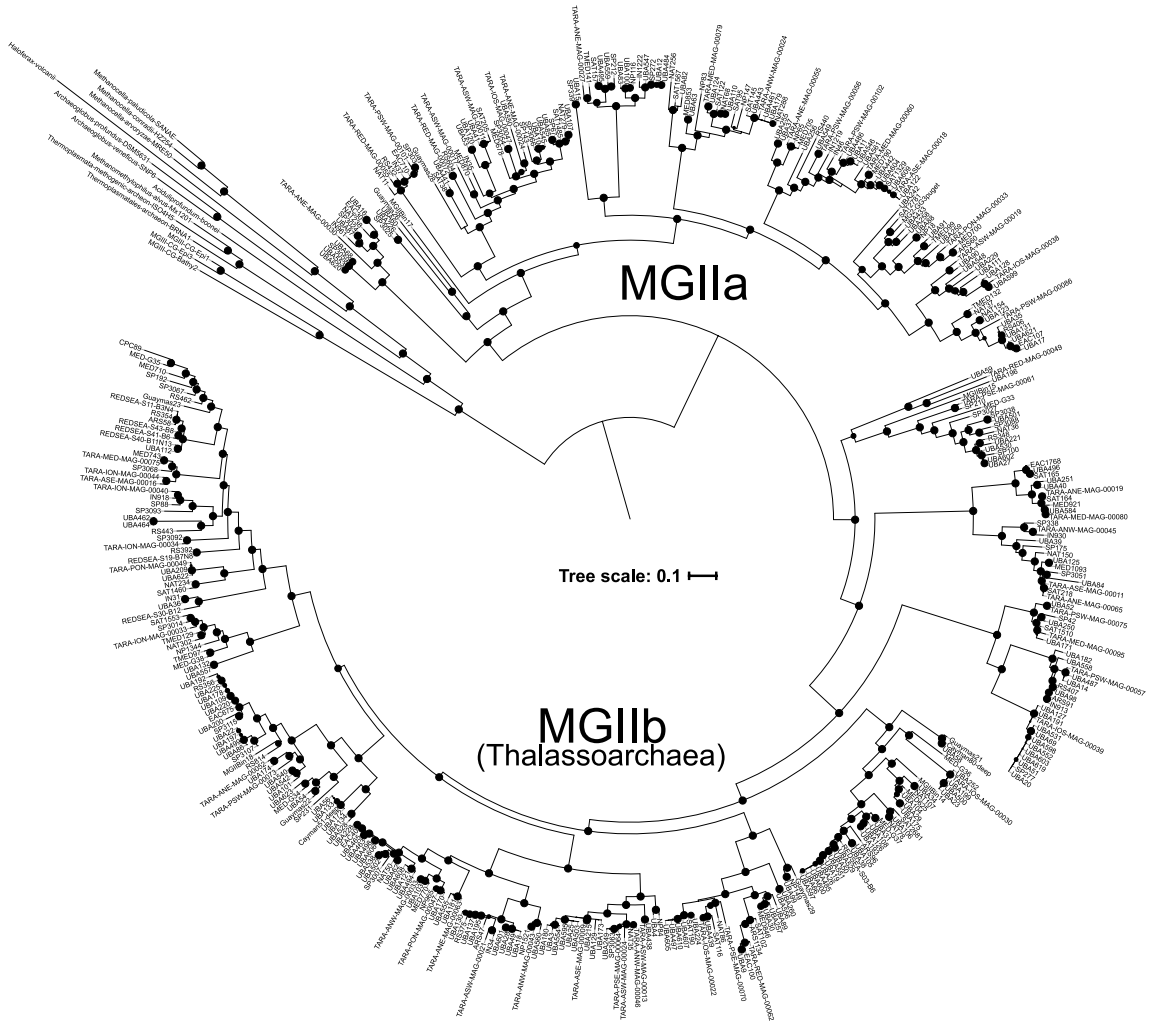
58. Boeuf, D., Audic, S., Brillet-Guéguen, L., Caron, C. & Jeanthon, C. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015,** bav080–8 (2015).
59. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* **5,** e11147 (2010).
60. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421–9 (2009).
61. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **6,** e4320–23 (2018).
62. 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).
63. van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804,** 281–295 (2012).
64. Kopf, A. *et al.* The ocean sampling day consortium. *GigaScience* **4,** 27 (2015).
65. Li, H., Handsaker, B., Fennell, T., Ruan, J. & Homer, N. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
66. Legendre, P. & Legendre, L. *Numerical Ecology*. **24,** (Elsevier, 1998).
67. Hammer, Ø., Harper, D. & Ryan, P. D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4,** 9 (2001).
68. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26,** 32–46 (2001).
69. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* **27,** 326–349 (1957).
70. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B Methodological* **57,** 289–300 (1995).

Supplemental Information

Supplemental Table 1. Information for all genomes used in study, including source ID numbers, clade and subclade assignment, completion stats (length, percent complete, percent contamination, percent strain heterogeneity), and source reference.
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 MGII', 'Core MGIIa', or 'Core MGIIb'. Only includes proteins with putative KEGG assignments.
Supplemental Table 6. Corresponding metadata (environmental parameters) for *Tara* Oceans and Ocean Sampling Day samples.

Supplemental Data 1. Groups of MGII genomes with ≥98.5% ANI denoting representative and duplicate genome IDs.
Supplemental Data 2. Spreadsheet of the pairwise ANI and AAI values for MGIIa and MGIIb.
Supplemental Data 3. FASTA format of the 120 marker proteins for all MGII genomes.
Supplemental Data 4. Newick file of the phylogenomic tree generated using 120 concatenated marker proteins for the full redundant set of MGII genomes with ≥60 markers. Corresponds to Supplemental Figure 1.

715 Supplemental Data 5. Newick file of the phylogenomic tree generated using 120 concatenated marker
716 proteins for the analyzed set of MGII genomes. Corresponds to Figure 1.
717 Supplemental Data 6. FASTA format of the 16S rRNA gene sequences present in the MGII genomes.
718 Supplemental Data 7. Newick file of the phylogenetic tree generated using the 16S rRNA gene sequences
719 present in the MGII genomes. Corresponds to Supplemental Figure 4.
720 Supplemental Data 8. Pfam HMMs used to identify MEROPS peptidases. The link between Pfams and
721 MEROPS can be found in Supplemental Table 4.
722 Supplemental Data 9. FASTA format of the putative proteorhodopsin sequences identified in the MGII
723 genomes.
724 Supplemental Data 10. Newick file of the phylogenetic tree generated using the proteorhodopsin
725 sequences present in the MGII genomes. Corresponds to Supplemental Figure 6.
726 Supplemental Data 11. GenBank files gathered from NCBI for the putative archaeal flagellum operon for
727 Tully *et al*. (2018) genomes.
728 Supplemental Data 12. Raw read counts for all MGII contigs generated from *Tara* Oceans and Ocean
729 Sampling Day samples.
730 Supplemental Data 13. Derived relative fraction and RPKM values for all MGII genomes for all *Tara*
731 Oceans and Ocean Sampling Day samples. Corresponds to Figure 5 and Supplemental Figure 8.
732 Supplemental Data 14. Spreadsheet of the PERMANOVA statistics (F statistic, *p*-value, and *p*-value
733 Benjamini-Hochberg False Discovery Rate) and the results of the ad hoc pairwise PERMANOVA tests
734 (*p*-value Bonferroni corrected).
735
736
737

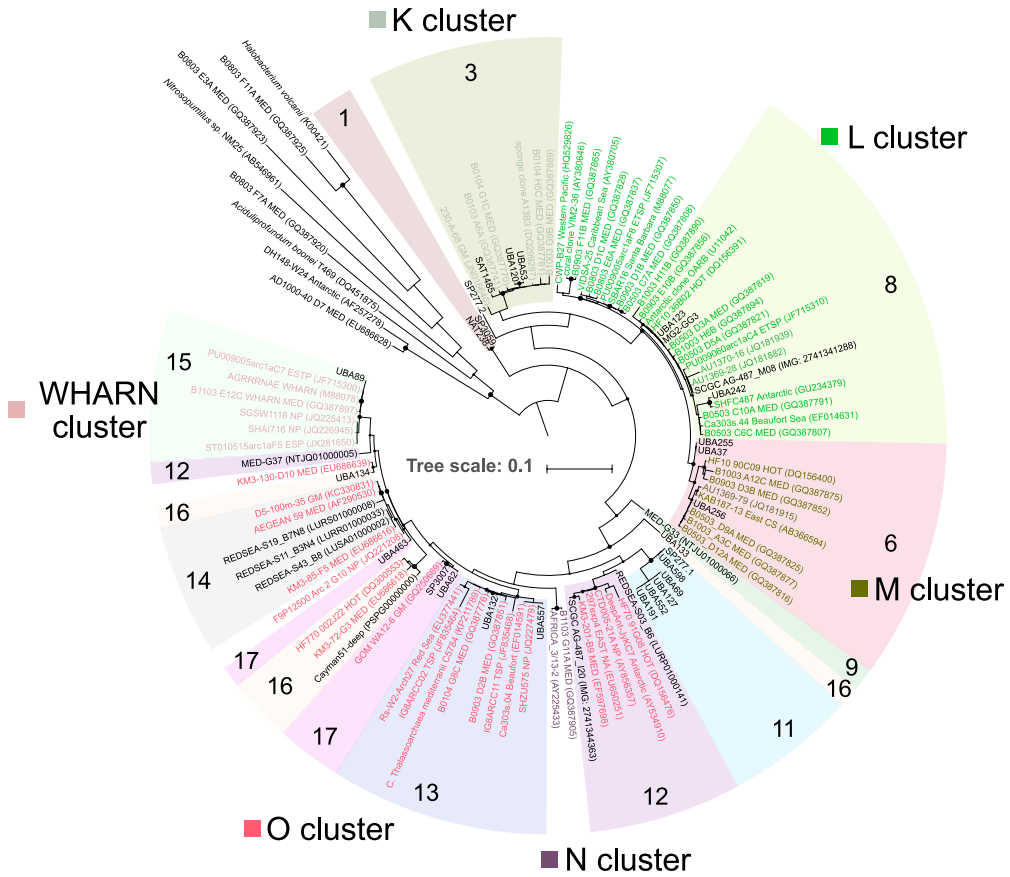


Supplemental Figure 1. A phylogenomic tree constructed using 120 concatenated marker proteins for all of the MGII. Bootstrap values are scaled proportionally between 0.75-1.

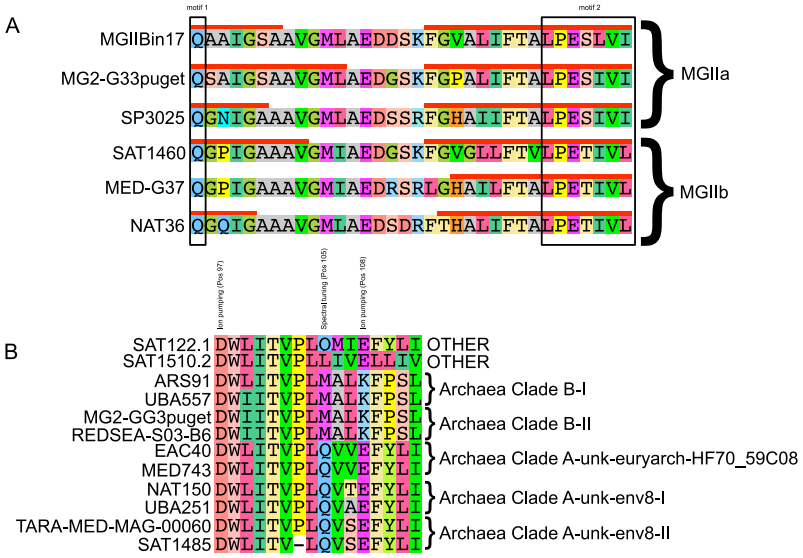
742 [AVAILABLE THROUGH FIGSHARE DUE TO SIZE]
743 Supplemental Figure 2. Heatmap of pairwise AAI and ANI for the MGIIa. Genomes are ordered based on
744 their placement in the phylogenetic tree in Figure 1.
745

[AVAILABLE THROUGH FIGSHARE DUE TO SIZE]

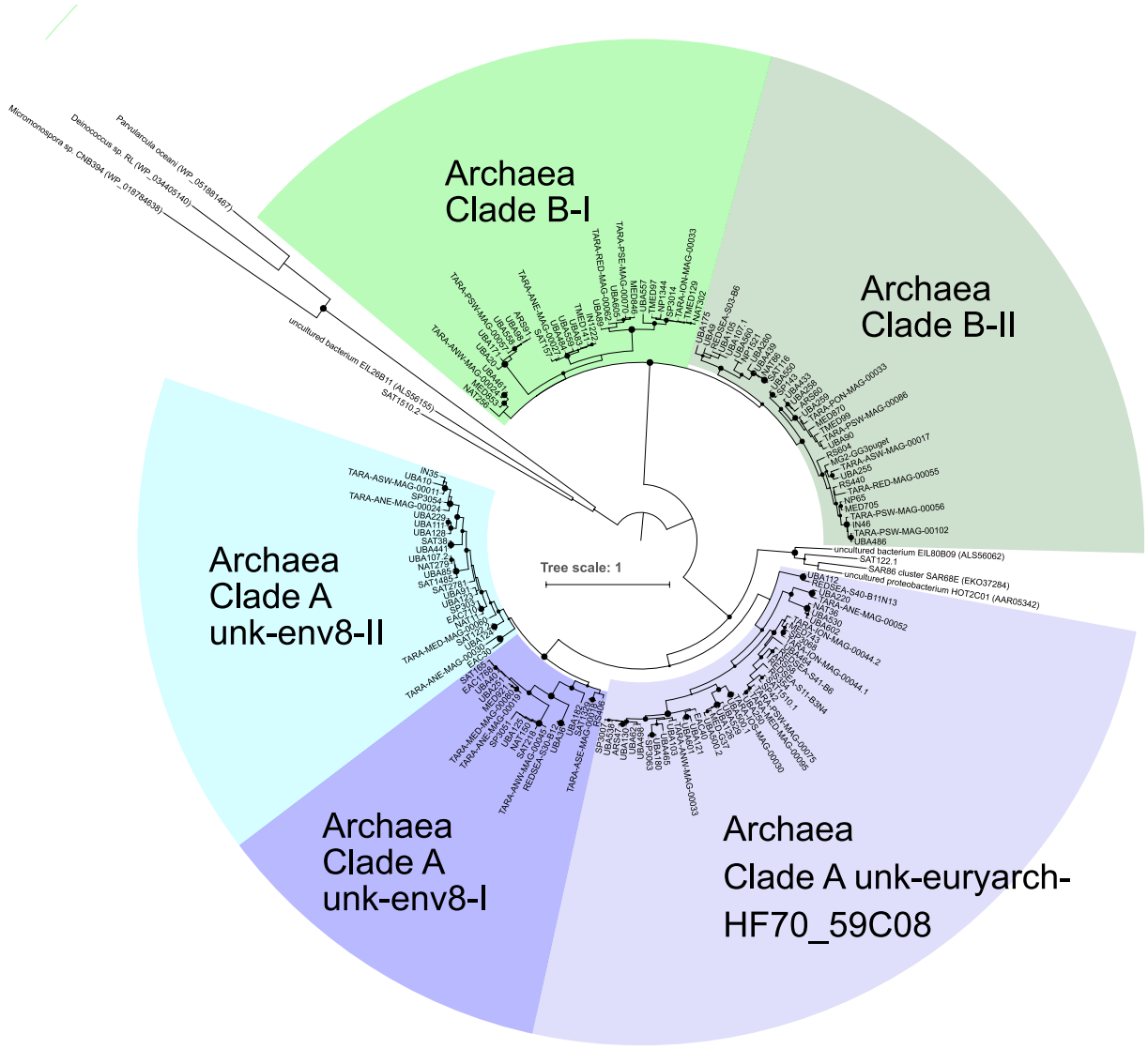
Supplemental Figure 3. Heatmap of pairwise AAI and ANI for the MGIIb. Genomes are ordered based on their placement in the phylogenetic tree in Figure 1.



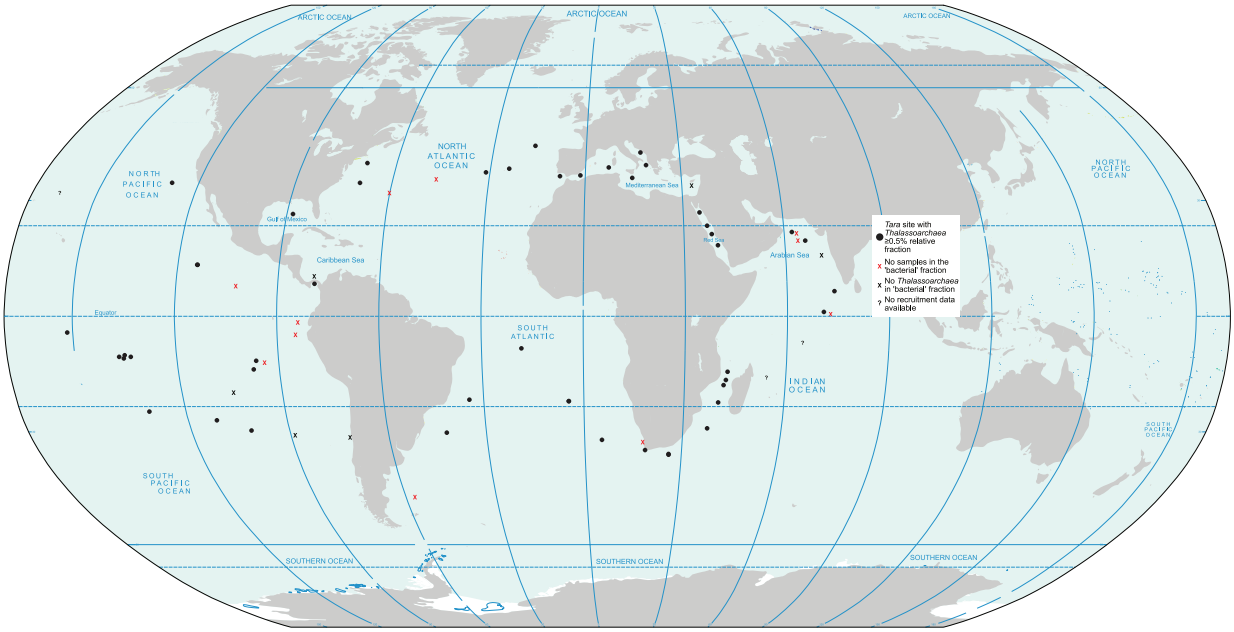
Supplemental Figure 4. A phylogenetic tree of the 16S rRNA gene for 35 MGII genomes combined with previously defined reference sequences. Previously observed clusters (leaves are colored to match corresponding clusters) that could be linked to newly defined subclades based on the occurrence of genome linked 16S rRNA sequences are shown. Internal nodes within large clusters (*e.g.*, O cluster) are used to demarcate different families where appropriate.



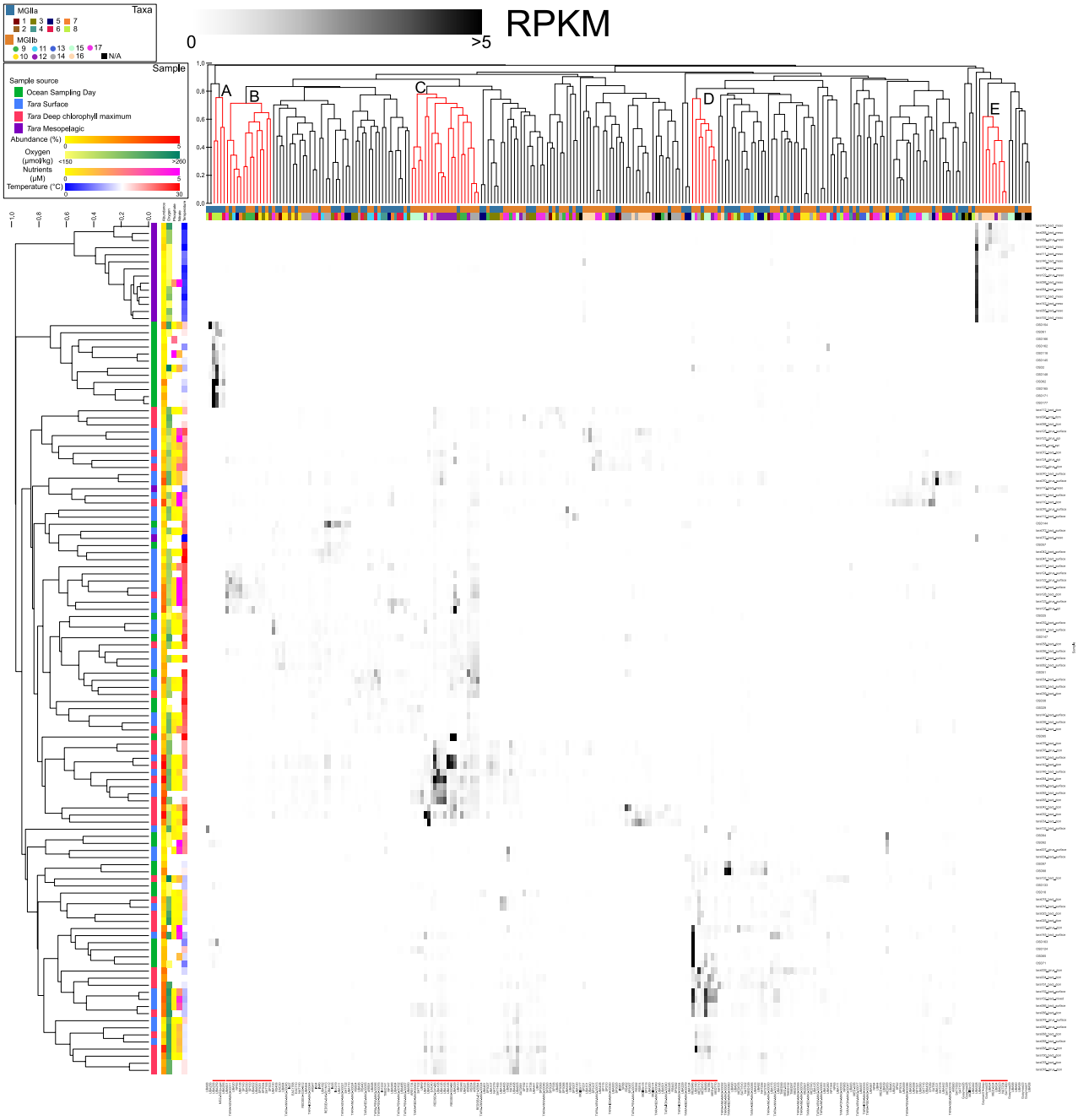
Supplemental Figure 5. A) Alignment *c* ring protein subunit (AtpK) for a selection of genomes. Transmembrane helices predicted using the TMHMM server (v.2.0) are denoted as red lines above the predicted region. Black boxes highlight the two conserved motifs that have been previously identified in Na^+ 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 6. "OTHER" refers to rhodopsin sequences that did not fall within Archaea Clade-A or -B.



Supplemental Figure 6. Phylogenetic tree of the rhodopsin predicted proteins (226AA alignment) identified in the MGII genomes. Subclades in previously established proteorhodopsin clades are denoted. Bootstrap values are scaled proportionally between 0.75-1.



Supplemental Figure 7. 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 8. A heatmap displaying the RPKM values for all of the MGII genomes in high abundance samples (≥0.5% relative fraction). RPKM values are scaled from 0-5 with values ≥5 in black (median, 0.0001; maximum, 9.25). Samples are average linkage hierarchically clustered based on a Bray-Curtis dissimilarity distance for all MGII RPKM values and the sample source is displayed. The available environmental parameters are presented as colored heatmaps (missing parameters are not displayed). The order of the sample hierarchical clustering and displayed environmental parameters are the same as those presented in Figure 5.

Supplementary Information

Supplemental Methods

As detailed in the Methods section, several genomes originating from Tully *et al.* (2017, 2018) were identified that had elevated estimated genome contamination levels (~5% or greater). To access the data of these refinement steps for the five genomes present in the analyzed genome dataset, the Anvi'o contigs (contigs.db) and profile (PROFILE.db) have been provided (figshare). The (iterative) manual binning results can be accessed using the Anvi'o command anvi-interactive (-p PROFILE.db -c contigs.db). Contigs related to the TOBG genomes MED853 and MED921 and the TMED genome TMED132 were combined with overlapping bins generated by CONCOCT and used to recruit the 20 *Tara* Oceans metagenomes originating from the Mediterranean Sea. The same approach was used for TOBG genomes SAT1485 and SP3063, except that reads were recruited only from the bacterial and girus fractions originating from the corresponding 25 South Atlantic and 24 South Pacific metagenomes, respectively.

Generally for each genome and profile pair, there are three bin collections that can be overlaid on the contig recruitment data: 'CONCOCT' presents predicted bins as determined by CONCOCT; 'Stage1' presents the contigs identified in the 'original' genome; and, 'Version1' presents the contigs in the 'refined' genome and the contigs that remain in the 'original'. 'Refined' contigs were selected based on the location of the majority of contigs in the original genome and contigs were removed that are not within that cluster. Contigs were added to the genome if they clustered on the hierarchical dendrogram with the majority of contigs and had complementary recruitment patterns and percent GC-content. Refined genomes were assessed using CheckM for estimated completeness and contamination. Refined genomes that did not see an improvement in completion or contamination were further refined, as is the case with SP3063. SP3063 has two bin collections, 'Version1' and 'Version2', where 'Version1' represents the first manual refinement attempt ('attempt1') and 'Version2' represents the 'refined' genome.