Villanueva, L., Schouten, S. & Sinninghe Damsté, J.S. (2015). Depth-related distribution of a key gene of the tetraether lipid biosynthetic pathway in marine Thaumarchaeota. *Environmental Microbiology*, 17(10), 3527–3539

Depth-related distribution of a key gene of the tetraether lipid biosynthetic pathway in marine Thaumarchaeota

Laura Villanueva*, Stefan Schouten,

and Jaap S. Sinninghe Damsté

NIOZ Royal Netherlands Institute for Sea Research. Department of Marine Organic Biogeochemistry. P.O. Box 59, NL-1790 AB Den Burg, The Netherlands. *Correspondence to L.V. e-mail: laura.villanueva@nioz.nl

Running title: Distribution of a thaumarchaeotal lipid enzyme

Keywords: ammonia monooxygenase (*amo*A), geranylgeranylglyceryl phosphate (GGGP) synthase, glycerol dialkyl glycerol tetraether (GDGT), membrane lipids, Thaumarchaeota.

Summary

The distribution of isoprenoid glycerol dialkyl glycerol tetraethers (GDGT) lipids synthesized by Thaumarchaeota have been shown to be temperature dependent in world oceans. Depth-related differences in the ammonia monooxygenase (*amo*A) of Thaumarchaeota have led to the classification of 'shallow ' and 'deep water' clusters, potentially affecting GDGT distributions. Here, we investigate if this classification is also reflected in a key gene of the thaumarchaeotal lipid biosynthetic pathway coding for geranylgeranylglyceryl phosphate (GGGP) synthase. We investigated metagenomic databases, suspended particulate matter and surface sediment of the Arabian Sea oxygen minimum zone (OMZ). These revealed significant differences in *amo*A and GGGP synthase between 'shallow' and 'deep water' Thaumarchaeota. Intriguingly, *amo*A and GGGP synthase sequences of benthic Thaumarchaeota clustered with the 'shallow water' rather than with 'deep water' Thaumarchaeota. This suggests that pressure and temperature are unlikely factors that drive the differentiation and suggest an important role of ammonia concentration which is higher in benthic and 'shallow water' niches. Analysis of the relative abundance of GDGTs in the Arabian Sea and in globally distributed surface sediments showed differences in GDGT distributions from subsurface to deep waters that may be explained by differences in the GGGP synthase, suggesting a genetic control on GDGT distributions.

Introduction

Thaumarchaeota were initially known as marine group I Archaea and considered as members of the crenarchaeotal phylum based on 16S ribosomal RNA gene phylogeny (DeLong, 1992; Fuhrman *et al*., 1992). However, subsequent studies using comparative genomics revealed that they form a separate and deep-branching phylum within the Archaea (Brochier-Armanet *et al*., 2008; Spang *et al*., 2010). Ecophysiological studies suggest that the thaumarchaeotal phylum is highly diversified and present in a wide variety of ecosystems (marine, freshwater, soil and hot environments; e.g. Erguder *et al*., 2009; Hatzenpichler, 2012). This novel phylum comprises ammonia-oxidizing archaea (AOA) but also environmental sequences representing microorganisms of unknown metabolism (Pester *et al*., 2011). Thaumarchaeota are often abundant (i.e. they are estimated to represent up to 20% of all picoplanktonic cells in the world ocean; Karner *et al*., 2001) and are important players of the global nitrogen and carbon cycles (Francis *et al*. 2005; Wuchter *et al*., 2006).

In most marine environments, pelagic thaumarchaeotal gene abundance increases rapidly with depth with maximum copy numbers of *amoA* and 16S rRNA genes near the base of the photic zone or in the transitional waters separating the epipelagic zone from the mesopelagic zone (200−500 m depth) (Church *et al*. 2010, Beman *et al*., 2012; Lam *et al*., 2007; Santoro *et al*., 2010; Pitcher *et al*., 2011a). Thaumarchaeotal genes have also been detected in much deeper meso- and bathypelagic waters (DeLong, 1992; Fuhrman *et al*., 1992; Karner *et al*., 2001; Wuchter *et al*., 2005; Lam *et al*., 2007; Mincer *et al*., 2007; Beman *et al*., 2008; Church *et al*., 2010; Santoro *et al*., 2010; Pitcher *et al*., 2011a; Schouten *et al*., 2012). It has been shown that AOA in marine waters can be subdivided into ‘shallow’ and ‘deep water’ clusters (also known as cluster A and B, respectively) based on differences in their *amo*A gene sequence (Francis *et al*.,

2005; Hallam *et al*., 2006; Mincer *et al*., 2007; Beman *et al*., 2008; Santoro *et al*., 2010; Hu *et al*., 2011, among others). This differentiation into 'shallow' and 'deep water' clusters has also been observed for other thaumarchaeotal metabolic genes such as *acc*A and *ure*C (e.g. Yakimov *et al*., 2011; Hu *et al*., 2011). Recently, Luo *et al*. (2014) detected genes coding for photolyase and catalase exclusively in members of the epipelagic 'shallow' clade, suggesting an adaptation of this population to reduce light-induced damage. Several studies have considered selective factors, such as competition with other microbial groups, oxygen concentration, depth, temperature, or latitude, to explain the distribution of different clades of Thaumarchaeota in the ocean (Hallam *et al*., 206; Mincer *et al*., 2007; Erguder *et al*., 2009; Pester *et al*., 2012; Cao *et al*., 2013). Sintes *et al*. (2012) recently hypothesized that the biogeographic and depth-related distribution of AOA clusters may be related to differences in ammonia availability.

Thaumarchaeota synthesize isoprenoid glycerol dialkyl glycerol tetraethers (GDGTs) with 0–4 cyclopentane moieties (Fig. S1), and the GDGT crenarchaeol, which contains a cyclohexane moiety in addition to four cyclopentane moieties (Schouten *et al*., 2000; Sinninghe Damsté *et al*., 2002), as their membrane lipids. In addition, they also synthesize a crenarchaeol regioisomer, which has been found in low relative abundance with respect to crenarchaeol in the thaumarchaeotal group 1.1a, but at higher relative abundances in the thaumarchaeotal group 1.1b (Sinninghe Damsté *et al*., 2012). Several studies have suggested that crenarchaeol is exclusively synthesized by Thaumarchaeota (Sinninghe Damsté *et al*., 2002; 2012; de la Torre *et al*., 2008; Schouten *et al*., 2008a; Pitcher *et al*., 2009 and 2011b), and can be used as a suitable marker to trace this group. The distribution of thaumarchaeotal GDGTs in the marine environment has been shown to be affected by temperature, i.e. with increasing temperature there is an increase in the relative abundance of cyclopentane-containing

GDGTs (Schouten *et al*., 2002; Wuchter *et al*., 2004, 2005). Based on this relationship, the TEX_{86} paleotemperature proxy was developed and calibrated using sea surface temperature (e.g. Schouten *et al*., 2002; Kim *et al*., 2010), since it is thought that GDGTs in marine sediments derive mostly from surface-derived thaumarchaeotal biomass (e.g. Wakeham *et al*., 2003). However, radiocarbon measurements of GDGTs in Bermuda Rise and Santa Monica Basin sediments suggested that a substantial part of the GDGTs in marine sediments are not derived from subsurface waters (< 200 m), and that archaeal production in the deeper water column may contribute to the pool of GDGTs in sediments (Pearson *et al*., 2001; Shah *et al*., 2008). A recent study by Taylor *et al*. (2013) noted an increase of the GDGT-2/GDGT-3 ratio in marine suspended particulate matter (SPM) and surface sediments with increasing water depth, suggesting a possible contribution of deep-water Archaea to the sedimentary GDGT distribution. Considering these studies, it is important to improve our knowledge on the ecophysiology and niche preference of 'deep water' Thaumarchaeota to assess their contribution to the sedimentary GDGT pool and potential impact on TEX_{86} paleothermometry.

In order to investigate if the segregation of AOA in 'shallow' and 'deep water' clusters explains the observed differences in GDGT distribution in shallow and deep marine waters, we targeted a key gene involved in the archaeal ether biosynthetic pathway, i.e. geranylgeranylglyceryl phosphate (GGGP) synthase. This key enzyme catalyzes the formation of an ether bond between isoprenyl diphosphate and glycerol-1-phosphate to form GGGP and is the first committed step towards ether membrane lipid synthesis (Koga and Morii, 2007; Matsumi *et al*., 2011). We examined the diversity of archaeal GGGP synthase and *amo*A in SPM samples throughout the Arabian Sea OMZ, as well as in a surface sediment located underneath the SPM sampling station at 3003 m

depth. For these samples, GDGT distributions have been previously reported (Pitcher *et al*., 2011a; Schouten *et al*., 2012; Lengger *et al*., 2012). We also assessed the diversity and distribution of the genes encoding archaeal GGGP synthase and *amo*A in a wide variety of metagenomes obtained from the marine water column globally. Our specific aims were to (i) survey the occurrence of GGGP synthase homologs at 'shallow' and 'deep water' depths, and (ii) to compare the distribution of GGGP synthase homologs with reported GDGT lipid distributions.

Results

Diversity of thaumarchaeotal ammonia monoxygenase in the marine water column

We amplified, cloned and sequenced the archaeal *amo*A gene from Arabian Sea SPM (Pitcher *et al*., 2011a) from two depths: 170 and 1050 m, representing the 'shallow' and 'deep water' niches, respectively. These depths represent two peaks of thaumarchaeotal 16S rRNA gene abundance, which are characterized by similar oxygen concentrations (4.8 and 4.6 μM, respectively; Pitcher *et al*., 2011a; Table S1). *Amo*A gene sequences recovered from 170 m depth were closely related to each other and with *amo*A gene sequences previously reported in 'shallow water' environments (0–200 m depth; 'Water column A' clade of Francis *et al*., 2005), as well as in sediments (Figure 1). *Amo*A gene sequences amplified from 1050 m depth are also closely related with each other and with *amo*A gene sequences previously assigned to the 'Water column B, deep water' *amo*A clade (Francis *et al*., 2005), but are clearly divergent from the *amo*A gene 'shallow water' clade (Figure 1).

To determine the differences in amino acid composition of the *amo*A protein of 'shallow' versus 'deep water' clusters, we aligned and compared two representative partial *amo*A gene coding sequences, which were recovered from the Arabian Sea at

depths 170 and 1050 m (Figure 2A), thaumarchaeotal *amo*A proteins from cultivated species, and environmental sequences from metagenomic databases defined here as representative of the 'shallow' (< 200 m depth; e.g. sequence with accession number ACG69579 detected at 100 m depth in the South China Sea), and 'deep water' clusters (> 1000 m; e.g. accession number ACF75441, detected at 2267 m in the Juan de Fuca Ridge). Several amino acid differences are observed between sequences from the *amo*A 'shallow' and 'deep water' clusters, of which 15 were unequivocally different (Figure 2A). In general, most of the divergent amino acid positions imply a change for an amino acid of the same nature between *amo*A sequences of the 'shallow' and 'deep water' cluster, e.g. replacement of alanine (A) by glycine (G) or serine (S) (all 'small' amino acid residues), or replacement of glutamine (Q) by asparagine (N) (both acidic amino acids, labeled in orange in Figure 2A). However, two amino acid positions showed a more substantial change in the amino acid type (positions 94 and 125, labeled in green in Figure 2A), i.e. from an aromatic 'bulky' amino acid (i.e. tyrosine, Y; phenylalanine, F) to a 'small' amino acid residue (i.e. valine, V; alanine, A; leucine, L) or vice versa.

In order to determine the distribution of the 'shallow' and 'deep water' clusters in different depth intervals of the ocean, archaeal *amo*A protein sequences from oceanic water column were extracted from metagenomic databases by using two representative sequences of the *amo*A 'shallow' and 'deep water' clusters as query sequences (i.e. putative *amoA* sequence accession number ACG69579, South China Sea, 100 m deep; Hu *et al*., 2011, and *amo*A sequence ACF75441, Juan de Fuca Ridge, 2267 m deep; Wang *et al*., 2009). In addition, *amo*A sequences detected in the South China Sea, Black Sea, central Mediterranean Sea, Central Pacific, Monterey Bay, Antarctic waters, North Pacific subtropical Gyre, North Atlantic, eastern south Pacific, and central California current were included in the analysis (Francis *et al*., 2005; Hallam *et al*., 2006; Lam *et*

168 *al*., 2007; Mincer *et al*., 2007; Agogue *et al*., 2008; Molina *et al*., 2010; Santoro *et al*.,
169 2010; Hu *et al*., 2011; Yakimov *et al*., 2011). This resulted in 1067 annotated archaeal
170 *amo*A sequences, which could all be assigned to either the 'shallow' or the 'deep water'
171 clusters according to their amino acid composition as described in Figure 2A.
172 Comparison with the water depth from which they were recovered showed that at 0−200
173 m depth almost all the sequences belonged to the 'shallow water' cluster (Figure 2B).
174 The 'deep water' *amo*A cluster becomes abundant from 500 m depth onwards,
175 representing approximately 80% of the *amo*A sequences in waters deeper than 1000 m.

176 *Diversity of thaumarchaeotal GGGP synthases in the marine water column*

177 To investigate the diversity and distribution of the thaumarchaeotal GGGP synthase-
178 coding gene, a fragment of the thaumarchaeotal GGGP synthase-coding gene was
179 amplified from SPM recovered from different depths (20, 170, 300, 450, 600, 1200,
180 2000 m depth) across the Arabian Sea OMZ (Pitcher *et al*., 2011a). Putative GGGP
181 synthase protein sequences obtained from the Arabian Sea SPM were included in a
182 phylogenetic analysis with GGGP synthases of thaumarchaeotal cultures. From the
183 putative GGGP synthase sequences recovered from shallow water depths of the Arabian
184 Sea SPM (0–170 m), 82% were closely related to the GGGP synthases of
185 thaumarchaeotal species isolated from shallow water environments (Figure 3).
186 Sequences recovered from intermediate depths (300–450 m deep) were distributed
187 throughout the phylogenetic tree (Figure 3), but the majority (78%) clustered with the
188 sequences recovered from deeper waters. Finally, 98% of the total reads of putative
189 GGGP synthase sequences recovered from 600 to 2000 m deep were concentrated in the
190 same cluster and were clearly divergent from the GGGP synthase sequences from
191 shallow waters (Figure 3).

To determine differences in amino acid compositions of the GGGP synthases from 'shallow' and 'deep waters', putative GGGP synthases recovered from the Arabian Sea SPM, GGGP synthases annotated in the genomes of (enrichment) cultures of various thaumarchaeotal species (Könneke *et al*., 2005; Mosier *et al*., 2012; Kim *et al*., 2011), and two fosmid sequences recovered from 4000 m depth in the Hawaii Ocean time series station ALOHA (Konstantinidis and DeLong, 2008) were aligned as shown in Figure 4A. The alignment is restricted to 76 amino acid residues between the 101–177 position of the GGGP synthase sequence of '*Ca.* Nitrosoarchaeum koreensis' (accession number ZP_08668912), representing 31% of the entire protein.

Remarkable changes were observed in the nature of the amino acid residues in this part of the putative GGGP synthase sequences from 'shallow'(< 200m) and 'deep' (> 1000 m) waters. For example, in the amino acid residue of position 121 of the 'shallow' sequences (indicated in green in Figure 4A), leucine (L) or alanine (A), 'small' hydrophobic amino acids, are substituted by arginine (R), an amino acid with a larger and basic (polar) residue, in the sequences recovered from the 'deep' water. Also, at amino acid position 144 (indicated in orange in Figure 4A), valine (V)/ isoleucine (I), 'small' hydrophobic amino acids, are detected in the sequences from shallow water, while the 'bulky' hydrophobic amino acid residue phenylalanine (F) is found in the sequences recovered from deep water.

We defined the 'shallow' and 'deep water' GGGP synthase clusters according to the amino acid changes listed in Figure 4A and performed an extensive metagenomic search of GGGP synthase homologues of both clusters. The representative sequence of the 'shallow water' GGGP synthase of '*Ca.* Nitrosoarchaeum limnia' (ZP_08257740), and the 'deep water' cluster of the uncultured Group I marine crenarchaeota HF4000APKG8I13 fosmid sequence (4000 m deep; ABZ09772.1) were used as query

sequences in the metagenome search. A total of 1236 sequences were recovered as homologues to the query sequences and they could all be assigned to either the 'shallow' or the 'deep water' GGGP synthase cluster according to their distinct amino acid composition (Figure 4A). Comparison with the water depth from which they were recovered showed that the GGGP synthase 'shallow water' cluster is dominating the shallow water depths, especially between 0–200 m depth (> 95%; Figure 4B). Water depths between 200–500 m seem to represent transitional niches in which the percentage of GGGP synthase 'deep water' cluster sequences starts to increase (9% on average). The 'deep water' cluster sequences are subsequently increasingly prominent with increasing depth, representing more than 60% of the total sequences in waters deeper than 2000 m.

AmoA and GGGP synthase of benthic Thaumarchaeota

*Amo*A and GGGP synthase gene sequences were amplified and sequenced from the surface sediment (upper 0.25 cm) of the station P3000 located at 3003 meters below sea level (mbss), which was located underneath the SPM sampling location. All *amo*A sequences recovered from the surface sediment at 3003 mbss were closely related to *amo*A protein sequences previously reported in 'shallow water' and marine sediments (Figure 1). All surface sediment *amo*A protein sequences had the same key amino acids found in the 'shallow water' *amo*A cluster sequences as indicated in Figure 2A, with the exception of the amino acid residue in position 73 (labeled in blue in Figure 2A) in which 27% of the surface sediment *amo*A sequences had a glycine (G) residue, characteristic for the 'deep water' cluster sequences, rather than the alanine (A) residue of the 'shallow water' sequences.

All putative GGGP synthases detected in the surface sediment of station P3000 clustered with the 'shallow water' GGGP synthases detected in the SPM but were more

closely related to the GGGP synthases of isolates of Thaumarchaeota (Figure 3). The alignment of GGGP synthase protein sequences in Figure 4A indicates that all GGGP synthases recovered from the surface sediment at 3003 mbss had the characteristic amino acid residue of 'shallow water' sequences in the four amino acid positions considered in our analysis.

The benthic Thaumarchaeota detected in the surface sediment at 3003 meter depth have been previously suggested to be active based on the high relative abundances of the labile membrane lipid hexose-phosphohexose crenarchaeol (which is expected to degrade fast in surface sediments with high oxygen concentrations; Lengger *et al*., 2012). The activity of this benthic population was confirmed here by the detection of *amo*A gene expression in the surface sediment of station P3000 (approx. 2×10^4 *amo*A mRNA copies g dry weight^{-1}; Table S1).

Discussion

Niche separation of 'shallow' and 'deep water' Thaumarchaeota

Our analysis of the ammonia monooxygenase (*amo*A) diversity in the Arabian Sea dataset, as well as the metagenomic analysis, confirms the existence of *amo*A 'shallow' and 'deep water' clusters in marine waters, as reported previously (Francis *et al*., 2005; Hallam *et al*., 2006; Mincer *et al*., 2007; Beman *et al*., 2008; Santoro *et al*., 2010; Hu *et al*., 2011). Our study now reveals that another important archaeal gene, i.e. the gene coding for a key enzyme in the archaeal ether lipid biosynthetic pathway, GGGP synthase, can also define thaumarchaeotal 'shallow' and 'deep water' clusters both in the Arabian Sea OMZ (Figure 3), as well as for oceans in general (Figure 4). Interestingly, the speciation in 'shallow' and 'deep water' clusters is not obvious from the phylogeny of the 16S rRNA gene. For example, the composition of the partial 16S

266 rRNA gene sequences in the Arabian Sea SPM reported by Schouten *et al*. (2012) did
267 not reveal a difference in thaumarchaeotal 16S rRNA gene-based phylogeny with depth.
268 This suggests that thaumarchaeotal populations of the 'shallow' and 'deep water'
269 clusters cannot be differentiated based on their 16S rRNA gene diversity but can be
270 differentiated on basis of functional genes like *amo*A (e.g Francis *et al*., 2005), *acc*A
271 and *ure*C (e.g. Yakimov *et al*., 2011) and, as shown here, the gene encoding GGGP
272 synthase.
273 Hu *et al*. (2011) suggested oxygen availability as a factor determining the
274 diversification of the thaumarchaeotal *amo*A and *acc*A genes. In the case of *amo*A gene,
275 sequences of the 'shallow' and 'deep water' clusters have been retrieved from water
276 with different oxygen concentrations: *amo*A 'shallow water' cluster A sequence have
277 been detected in oxygenated surface waters and in the suboxic zone (Francis *et al*.,
278 2005; Lam *et al*., 2007; Mincer *et al*., 2007; Agogue *et al*., 2008; Beman *et al*., 2008),
279 and 'deep water' cluster *amo*A gene sequences have been reported in suboxic and well-
280 oxygenated deep water (Francis *et al*., 2005; Mincer *et al*., 2007; Agogue *et al*., 2008).
281 In our study, *amoA* sequences of the 'shallow' and 'deep water' clusters were found in
282 SPM recovered at different depths with essentially the same oxygen concentration (4.8
283 and 4.6 μM at 170 and 1050 m depth, respectively). In much the same way, putative
284 GGGP synthases of the 'shallow' and 'deep water' clusters were also detected in
285 Arabian Sea SPM with similar oxygen concentration (i.e. 170 m, 4.8 μM oxygen; 600
286 m, 3.3 μM oxygen), but also GGGP synthases of the 'deep' water cluster were present
287 at depths where the water has a higher oxygen concentration (i.e. 2000 m, 63 μM) than
288 those dominated by sequences of the 'shallow water' cluster (i.e. 170 m, 4.8 μM
289 oxygen). This suggests that the segregation of thaumarchaeotal GGGP synthase and

290 *amo*A-based clusters in shallow and deep waters is not related to the oxygen
291 concentration.
292 Other factors which may explain the Thaumarchaeota 'shallow' and 'deep water'
293 cluster segregation are environmental variables directly related to depth itself, such as
294 hydrostatic pressure and temperature. Global surveys of archaeal *amo*A gene sequences
295 suggest that depth but also latitude/location and temperature can explain most of the
296 *amo*A gene sequence variation (Biller *et al*., 2012; Peng *et al*., 2013). However, both
297 the GGGP synthase and ammonia monooxygenase sequences derived from
298 metagenomes analyzed in this study were obtained from diverse ocean sites from
299 different latitudes and a wide range of temperature regimes (Tables S2 and S3).
300 Importantly, both *amo*A and GGGP synthase sequences of the active population of
301 benthic Thaumarchaeota present in the surface sediment at 3003 m water depth were
302 closely related to sequences of the 'shallow water' cluster and different from those of
303 the 'deep water' cluster despite experiencing similar temperatures (1.4°C *vs* 3.2°C) and
304 hydrostatic pressures as the Thaumarchaeota of the 'deep water' clade. This suggests
305 that neither temperature nor pressure are key factors driving the diversification of the
306 two different populations of Thaumarchaeota in marine settings.
307 Another important factor which may influence the niches of the different clades of
308 Thaumarchaeota are nutrient concentrations. However, nitrate, nitrite and phosphate
309 concentrations were relatively stable over depth (Table S1), and thus seem not explain
310 the niche separation between 'shallow' and 'deep' water Thaumarchaeota. Ammonia
311 availability has recently been suggested to drive the *amo*A diversification of AOA in the
312 water column (Sintes *et al*., 2013). In our dataset, ammonia concentrations were similar
313 in 'shallow' and 'deep waters' (0.196 and 0.209 μM, at 170 and 1050 m, respectively)
314 but it should be noticed that these values may not reflect the actual turn-over rates of

315 ammonia. Indeed, gene expression of the *amo*A gene was higher in the shallow waters
316 compared to deeper water in the Arabian Sea (Pitcher *et al*., 2011a), which suggests a
317 lower ammonia oxidation activity of the deep water population of Thaumarchaeota.
318 Ammonia as the discriminating factor could also explain the clustering of the *amo*A
319 gene sequences in deep water surface sediment with the 'shallow water' homologues as
320 there is a high availability of ammonia in both shallow waters as well as sediment pore
321 waters due to the breakdown of organic matter (Table S1).
322 At this stage, we can only speculate that perhaps the difference in the amino acid
323 sequence of the *amo*A protein (Figure 2) of the 'deep water' clade compared to the one
324 of the 'shallow water' clade may result in a competitive advantage at relatively low
325 ammonium concentrations. Alternatively, these amino acid differences could be
326 functionally neutral, if ammonia availability is not a strong selective pressure driving
327 divergence between the two clades. In this respect it is noteworthy that Santoro *et al*.
328 (2010) did found 'deep water' *amo*A in shallow waters but no expression of the 'deep
329 water' *amo*A gene was detected, suggesting they are inhibited once transported to the
330 surface. Taken together this may suggest that, possibly driven by ammonia availability,
331 two thaumarchaeotal populations evolved divergently from a common ancestor and
332 adapted to specific conditions, leading to differences not only in the *amo*A gene but also
333 other genes like the one coding for GGGP synthases.

334 *Potential impact of different GGGP synthases on GDGT lipid distribution*

335 To examine whether the differences in the GGGP synthase gene in the 'shallow' *vs*
336 the 'deep water' populations of Thaumarchaeota (Figure 4) may have an impact on
337 GDGT composition, we compared the thaumarchaeotal GGGP synthase distribution
338 with the distribution of GDGTs in the Arabian Sea SPM as reported by Schouten *et al*.
339 (2012). The fractional abundances of GDGTs change with depth for both intact polar

lipids (IPLs; Figure S2A), representing living biomass, as well as for core lipids (CLs; Figure S2B), representing dead material. The most striking difference is an increased fractional abundance of IPL-GDGT-2 with depth with a concomitant drop in the fractional abundance of IPL-GDGT-3 with depth (Figure S2A, B). Examination of the relative abundances of the CL-GDGTs in the Equatorial Pacific (EqPac) station 6 during two campaigns (February–March 1992; August–September 1992; Wuchter *et al*., 2005) revealed a similar pattern, i.e. a steady increase in CL-GDGT-2 and a decrease of CL-GDGT-3 relative abundances with depth (Figure S2C, D), as previously indicated by Taylor *et al*., (2013) and more recently by Hernandez-Sanchez *et al*. (2014) in SPM and underlying sediments collected in the Southeast Atlantic Ocean. Generally, GDGT-2 and -3 in marine waters are assumed to be both derived from Thaumarchaeota (Schouten *et al*., 2013) but the possibility that the uncultured marine euryarchaeota group II synthesize them has been previously raised and debated (Turich *et al*., 2007; Schouten *et al*., 2008b). However in our data set, as shown in Figure S3, compiling data by Pitcher *et al*., (2011a) and Schouten *et al*., (2012), the general archaeal 16S rRNA gene quantification in the Arabian Sea SPM almost matched the thaumarchaeotal 16S rRNA gene quantification, supporting that most Archaea present in the water column were members of the Thaumarchaeota and thus likely to be the dominant GDGT producers in the system.

In the water column of the Arabian Sea OMZ, the GDGT-2/GDGT-3 ratio increased with depth reaching a maximum at 1200 m depth (Figure 5A, Figure S4). Interestingly, this change in the relative abundance of GDGT with depth coincides with the depth niches occupied by ‘shallow’ and ‘deep water’ Thaumarchaeota (Figure 5B), tentatively indicating a potential link between these clusters and the differences in GDGT distributions. The same trend is observed globally with an increased GDGT-2/GDGT-3

ratio in surface sediments with depth (Figure 5C; based on data reported by Kim *et al*., 2010), and the increasing percentage of GGGP synthases of the 'deep water' cluster in the marine metagenomes (Figure 5D; Figure 4B). Villanueva *et al.* (in press) suggested, based on previously reported experimental evidence and phylogenetic analyses of the enzymes involved in the GDGT pathway, that the ring moieties present in GDGTs may already be present in the isoprenyl substrates processed by the GGGP synthase, thus suggesting that the GGGP synthase is already involved in determining the relative abundance of the different GDGTs synthesized in this pathway. Considering this hypothesis, differences in the amino acid sequences of GGGP synthases between different archaeal groups could result in differences in the relative proportion of the GDGTs. In the present study, we have unraveled important changes in the amino acid composition of the catalytic site of the GGGP synthase (labeled in purple and green in Figure 4A) present in the thaumarchaeotal populations inhabiting 'shallow' and 'deep waters'. We hypothesize that those differences may lead to differences in the relative proportion of synthesized GDGTs, e.g. a strongly different ratio of GDGT-2 over GDGT-3. Similarly, it has been found that Thaumarchaeota of the group I.1b have a different GDGT composition than those of I.1a, i.e. a higher abundance of the crenarchaeol regio-isomer (Sinninghe Damste *et al*., 2012). Thus, with increasing depth the increasing dominance of the 'deep water' clade of Thaumarchaeota may cause the observed increase in GDGT-2 over GDGT-3 (Figure 5). This would also explain why the GDGT distributions apparently do not change with lower temperatures in the deep sea (Wuchter *et al*., 2005; Schouten *et al*., 2012), i.e. the GDGT-3 would be expected to decrease compared to GDGT-2 in Thaumarchaeota with lower temperatures (Wuchter *et al*., 2004; Schouten *et al*., 2007) . It is also possible that other genes in the pathway, not just GGGP synthase, are responsible for the observed variation in the GDGT ratios with

depth, or that regulatory genes play a role. Biochemical experiments with GGGP synthase variants, as well as cultivation of AOA isolated from different depths will be required to corroborate this hypothesis

Experimental procedures

DNA/RNA extraction and quantitative PCR

DNA/RNA was extracted from surface sediment (upper 0.25 cm) with the RNA PowerSoil® Total Isolation kit plus the DNA elution accessory (Mo Bio Laboratories, Inc., Carlsbad, CA). RNA extracts were treated with DNAse and reverse transcribed as described by Pitcher *et al*. (2011a). Quantitative PCR of the *amo*A gene in surface sediments was performed as described by Pitcher *et al*. (2011a).

Cloning, sequencing and phylogeny of the archaeal amoA gene

Amplification of the archaeal *amo*A gene was performed as described by Yakimov *et al*. (2011). PCR reaction mixture was the following (final concentration): Q-solution 1× (PCR additive, Qiagen); PCR buffer 1×; BSA (200 μg ml^{-1}); dNTPs (20 μM); primers (0.2 pmol μl^{-1}); $MgCl_2$ (1.5 mM); 1.25 U Taq polymerase (Qiagen, Valencia, CA, USA). PCR conditions for these amplifications were the following: 95°C, 5 min; 35 × [95°C, 1 min; 55°C, 1 min; 72°C, 1 min]; final extension 72°C, 5 min. PCR products were gel purified (QIAquick gel purification kit, Qiagen) and cloned in the TOPO-TA cloning® kit from Invitrogen (Carlsbad, CA, USA) and transformed in *E. coli* TOP10 cells following the manufacturer's recommendations. Recombinant clones plasmid DNAs were purified by Qiagen Miniprep kit and screening by sequencing ($n \geq 30$) using M13R primer by Macrogen Europe Inc. (Amsterdam, The Netherlands). Obtained archaeal *amo*A protein sequences were aligned with already annotated *amo*A sequences by using the Muscle application (Edgar, 2004). Phylogenetic trees were constructed with the Neighbor-Joining method (Saitou and Nei, 1987) and evolutionary distances computed using the Poisson correction method with a bootstrap test of 1,000 replicates.

418 *Metagenomic search*

419 GGGP synthase protein sequences of Thaumarchaeota were obtained by protein blast
420 (pBLAST) using the annotated geranylgeranylglyceryl phosphate synthase protein
421 sequence of *Nitrosopumilus maritimus* (YP_001583129) as a query sequence (e-value ≤
422 $1e^{-109}$ and identity ≥ 73%): '*Ca*. Nitrosoarchaeum koreensis (YP_006774700),
423 *Cenarchaeum symbiosum* (YP_876742), '*Ca*. Nitrosopumilus salaria' (ZP_10117002),
424 '*Ca.* Nitrosoarchaeum limnia' (ZP_08257740), and '*Ca*. Nitrososphaera gargensis'
425 (YP_006864169). In addition, two environmental sequences (e-value ≤ $1e^{-101}$ and
426 identity ≥ 64%) annotated as PcrB family proteins (ABZ09772.1 and ABZ07221.1) in
427 the uncultured Group I marine crenarchaea HF4000APKG8I13 (EU016657.1) and
428 HF4000ANIW133C7 (EU016595.1) fosmid sequences were annotated as putative
429 GGGP synthases. GGGP synthase sequences of '*Ca.* Nitrosoarchaeum limnia'
430 (ZP_08257740), putative GGGP synthase (ABZ09772.1), putative ammonia
431 monooxygenase subunit A (*amo*A) ACF75441.1 (Juan de Fuca Ridge, 2267 m deep;
432 Wang *et al*., 2009), and ACG69579.1 (South China Sea, 100 m deep; Hu *et al*., 2011)
433 were used as query sequences in pBLAST with an e-value of $1e^{-20}$ in the Integrated
434 Microbial Genomes (IMG) system with microbiome samples (IMG/M) of the Joint
435 Genome Institute U.S Department of Energy (DOE) (http://img.jgi.doe.gov). The
436 pBLAST was restricted to search homologues in marine environmental microbiome
437 metagenomes listed in Table S2. Protein blast searches were also performed in NCBI
438 metagenomic environmental proteins database, and in the Community
439 Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis portal
440 (https://portal.camera.calit2.net) against the GOS (Global Ocean Sampling
441 Expedition) and HOT (Hawaii Ocean Time-series) all ORF peptides databases with
442 restricted e-value $1e^{-20}$. Tblastn (search translated nucleotide databases using a protein

443 query) searches were also performed in the Camera 2.0 portal against all metagenomic
444 454 reads and a custom dataset comprising deep-water samples (500 to 4857 m depth)
445 as listed in Table S3. GGGP synthase and ammonia monooxygenase protein sequence
446 homologues obtained by the metagenomic search were aligned with the Muscle
447 alignment application (Edgar, 2004) included in the Mega5 software (Tamura *et al*.,
448 2011).

449 *GGGP synthase gene primer design, sequences and phylogeny*
450 Primer pairs for the amplification of a part of GGGP synthase coding gene in
451 environmental samples were designed manually based on the alignment of GGGP
452 synthase genes of thaumarchaeotal cultures (as indicated in Figure 4A), as well as
453 metagenomic sequences. Sequences were aligned by Muscle (Edgar, 2004) and edited
454 manually. Primers were checked for secondary structures and % G+C in the Primer3
455 webpage (http://primer3.wi.mit.edu/) (see Table S4). PCR reactions and conditions were
456 performed as specified above. A gradient PCR cycle from 45 to 60°C melting
457 temperature was performed for each set of forward and reverse primers (Table S4) with
458 genomic DNA samples obtained from SPM recovered at different depths (20, 170, 300,
459 450, 600, 1200, 2000 m depth) across the Arabian Sea oxygen minimum zone during a
460 campaign in January 2009 (see Pitcher *et al*., 2011a for details on sampling conditions
461 and DNA extraction procedure) and genomic DNA from *Nitrosopumilus maritimus*
462 SCM1 as a positive control). The only primer pair that gave a positive PCR
463 amplification was GGGP_Thaum_301F_short/530R_short, as listed in Table S4.
464 Positive amplification bands were excised from agarose gel and gel or PCR purified
465 (QIAquick gel/PCR purification kit, Qiagen) and cloned in the TOPO-TA cloning® kit
466 from Invitrogen (Carlsbad, CA, USA) and transformed in *E. coli* TOP10 cells following

the manufacturer's recommendations. Recombinant plasmidic DNA was sequenced as described above.

Putative partial GGGP synthase gene sequences obtained from the Arabian Sea SPM samples were translated to protein by submitting them as query sequences in translated blast (xblast: Find similar proteins to translated query in a protein database) and reviewed by manual annotation. Putative and annotated partial GGGP synthases sequences were translated to protein and aligned by Muscle (Edgar, 2004) in Mega5 software (Tamura *et al*., 2011) and edited manually. Phylogenetic reconstruction of putative partial GGGP synthase proteins was performed by maximum likelihood in PhyML v3.0 (Guindon *et al*., 2010) using the LG model plus gamma distribution (LG+G) indicated by ProtTest 2.4 (Abascal *et al*., 2005). Branch support was calculated with the approximate likelihood ratio test (aLRT).

Nucleotide accession numbers

Partial GGGP synthase and *amo*A gene sequences sequence data have been submitted to the GenBank database under accession No: KF512026−KF512324 and KJ509833−KJ509868 (GGGP synthase), KF512325−KF512379 and KJ509763−KJ509795 (*amo*A).

Acknowledgments

We thank the captain and crew of the R/V Pelagia for their assistance in collecting SPM and sediments from the Arabian Sea. Dr. Gert-Jan Reichart (Utrecht University) for leading the PASOM cruise and Dr. Angela Pitcher and Dr. Sabine Lengger for providing the samples. We thank Elda Panoto for technical assistance.

References

Abascal, F., Zardoya, R., and Posada, D. (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21: 2104-2105.

Agogue, H., Brink, M., Dinasquet, J., and Herndl, G.J. (2008) Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. Nature 456: 788-791.

Beman, J.M., Popp, B.N., and Francis, C.A. (2008) Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. ISME J 2: 429-441.

Beman, J.M., Popp, B.N., and Alford, S.E. (2012) Quantification of ammonia oxidation rates and ammonia-oxidizing archaea and bacteria at high resolution in the Gulf of California and eastern tropical North Pacific Ocean. Limnol Oceanog 57: 711-726.

Biller, S.J., Mosier, A.C., Wells, G.F., and Francis, C.A. (2012) Global Biodiversity of Aquatic Ammonia-Oxidizing Archaea is Partitioned by Habitat. Front Microbiol 3: 252.

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

Cao, H., Auguet, J.C., and Gu, J.D. (2013) Global ecological pattern of ammonia-oxidizing archaea. Plos One 8: e52853.

Church, M.J., Wai, B., Karl, D.M., and DeLong, E.F. (2010) Abundances of crenarchaeal amoA genes and transcripts in the Pacific Ocean. Environ Microbiol 12: 679-688.

de la Torre, J.R., Walker, C.B., Ingalls, A.E., Konneke, M., and Stahl, D.A. (2008) Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. Env Microbiol 10: 810-818.

DeLong, E.F. (1992) Archaea in coastal marine environments. Proc Nat Acad Sci USA 89: 5685-5689.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinform 5: 113.

Erguder, T.H., Boon, N., Wittebolle, L., Marzorati, M., and Verstraete, W. (2009) Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. FEMS Microbiol Rev 33: 855-869.

Francis, C.A., Roberts, K.J., Beman, J.M., Santoro, A.E., and Oakley, B.B. (2005) Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Nat Acad Sci USA 102: 14683-14688.

525 Fuhrman, J.A., McCallum, K., and Davis, A.A. (1992) Novel major archaebacterial
526 group from marine plankton. Nature 356: 148-149.

527 Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O.
528 (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies:
529 Assessing the Performance of PhyML 3.0. Syst Biol 59: 307-321.

530 Hallam, S.J., Mincer, T.J., Schleper, C., Preston, C.M., Roberts, K., Richardson, P.M. et
531 al (2006) Pathways of carbon assimilation and ammonia oxidation suggested by
532 environmental genomic analyses of marine Crenarchaeota. Plos Biology 4: e95.

533 Hatzenpichler, R (2012) Diversity, physiology, and niche differentiation of ammonia-
534 oxidizing archaea. Appl Environ Microbiol 78: 7501-7510.

535 Hernandez-Sanchez, M.T., Woodward, E.M.S., Taylor, K.W.R., Henderson, G.M., and
536 Pancost, R.D. (2014) Variations in GDGT distributions through the water column in the
537 South East Atlantic Ocean. Geochim Cosmochim Acta 132:337-348.

538 Hu, A., Jiao, N., Zhang, R., and Yang, Z. (2011) Niche partitioning of marine group I
539 Crenarchaeota in the euphotic and upper mesopelagic zones of the East China Sea. Appl
540 Environ Microbiol 77: 7469-7478.

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

543 Kim, J.-H., van der Meer, M., Schouten, S., Helmke, P., Willmott, V., Sangiorgi, F. et al
544 (2010) New indices and calibrations derived from the distribution of crenarchaeal
545 isoprenoid tetraether lipids: Implications for past sea surface temperature
546 reconstructions. Geochim Cosmochim Acta 74: 4639-4654.

547 Kim, B.K., Jung, M.Y., Yu, D.S., Park, S.J., Oh, T.K., Rhee, S.K. et al (2011) Genome
548 sequence of an ammonia-oxidizing soil archaeon, "Candidatus Nitrosoarchaeum
549 koreensis" MY1. J Bacteriol 193: 5539-5540.

550 Koga, Y., and Morii, H. (2007) Biosynthesis of ether-type polar lipids in archaea and
551 evolutionary considerations. Microbiol Mol Biol Rev 71: 97-120.

552 Könneke, M., Bernhard, A.E., de la Torre, J.R., Walker, C.B., Waterbury, J.B., and
553 Stahl, D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon.
554 Nature 437: 543-546.

555 Konstantinidis, K.T., and DeLong, E.F. (2008) Genomic patterns of recombination,
556 clonal divergence and environment in marine microbial populations. ISME J 2: 1052-
557 1065.

558 Lam, P., Jensen, M.M., Lavik, G., McGinnis, D.F., Muller, B., Schubert, C.J. et al
559 (2007) Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea.
560 Proc Nat Acad Sci USA 104: 7104-7109.

561 Lengger, S.K., Hopmans, E.C., Reichart, G.-J., Nierop, K.G.J., Sinninghe Damsté, J.S.,
562 and Schouten, S. (2012) Intact polar and core glycerol dibiphytanyl glycerol tetraether
563 lipids in the Arabian Sea oxygen minimum zone. Part II: Selective preservation and

564 degradation in sediments and consequences for the TEX_{86}. Geochim Cosmochim Acta
565 98: 244-258.

566 Luo, H., Tolar, B.B., Swan, B.K., Zhang, C.L., Stenauskas, R., Moran, M.A., and
567 Hollibaugh, J.T. (2014) Single-cell genomics shedding light on marine Thaumarchaeota
568 diversification. ISME J 8: 732-736.

569 Matsumi, R., Atomi, H., Driessen, A.J., and van der Oost, J. (2011) Isoprenoid
570 biosynthesis in Archaea-biochemical and evolutionary implications. Res Microbiol
571 162:39-52.

572 Mincer, T.J., Church, M.J., Taylor, L.T., Preston, C., Karl, D.M., and DeLong, E.F.
573 (2007) Quantitative distribution of presumptive archaeal and bacterial nitrifiers in
574 Monterey Bay and the North Pacific Subtropical Gyre. Environ Microbiol 9: 1162-
575 1175.

576 Molina, V., Belmar, L., and Ulloa, O. (2010) High diversity of ammonia-oxidizing
577 archaea in permanent and seasonal oxygen-deficient waters of the eastern South Pacific.
578 Environ Microbiol 12: 2450-2465.

579 Mosier AC, Lund MB, Francis CA (2012) Ecophysiology of an ammonia-oxidizing
580 archaeon adapted to low-salinity habitats. Microb Ecol 64: 955-963.

581 Pearson, A., McNichol, A.P., Benitez-Nelson, B.C., Hayes, J.M., and Eglinton, T.I.
582 (2001) Origins of lipid biomarkers in Santa Monica Basin surface sediment: a case
583 study using compound-specific $\Delta^{14}C$ analysis. Geochim Cosmochim Acta 65: 3123-
584 3137.

585 Peng, X., Jayakumar, A., and Ward, B.B. (2013) Community composition of ammonia-
586 oxidizing archaea from surface and anoxic depths of oceanic oxygen minimum zones.
587 Front Microbiol 4: 177.

588 Pester, M., Schleper, C., and Wagner, M. (2011) The Thaumarchaeota: an emerging
589 view of their phylogeny and ecophysiology. Curr Opin Microbiol 14: 300-306.

590 Pester, M., Rattei, T., Flechl, S., Grongroft, A., Richter, A., Overmann, J. et al (2012)
591 amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing
592 of *amo*A genes from soils of four different geographic regions. Environ Microbiol 14:
593 525-539.

594 Pitcher, A., Rychlik, N., Hopmans, E.C., Spieck, E., Rijpstra, W.I., Ossebaar, J.,
595 Schouten, S., Wagner, M., and Sinninghe Damsté, J.S. (2009) Crenarchaeol dominates
596 the membrane lipids of Candidatus Nitrososphaera gargensis, a thermophilic Group I.1b
597 Archaeon. ISME J 4: 542-552.

598 Pitcher, A., Villanueva, L., Hopmans, E.C., Schouten, S., Reichart, G.J., and Sinninghe
599 Damsté, J.S. (2011a) Niche segregation of ammonia-oxidizing archaea and anammox
600 bacteria in the Arabian Sea oxygen minimum zone. ISME J 5: 1896-1904.

601 Pitcher, A., Hopmans, E.C., Mosier, A.C., Park, S.-J., Rhee, S.-K., Francis, C.A. et al
602 (2011b) Core and Intact Polar Glycerol Dibiphytanyl Glycerol Tetraether Lipids of

603 Ammonia-Oxidizing Archaea Enriched from Marine and Estuarine Sediments. Appl
604 Environ Microbiol 77: 3468-3477.

605 Saitou, N., and Nei, M. (1987) The neighbor-joining method: a new method for
606 reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

607 Santoro, A.E., Casciotti, K.L., and Francis, C.A. (2010) Activity, abundance and
608 diversity of nitrifying archaea and bacteria in the central California Current. Environ
609 Microbiol 12: 1989-2006.

610 Schouten, S., Hopmans, E.C., Pancost, R.D., and Sinninghe Damsté, J.S. (2000)
611 Widespread occurrence of structurally diverse tetraether membrane lipids: Evidence for
612 the ubiquitous presence of low-temperature relatives of hyperthermophiles. Proc Nat
613 Acad Sci USA 97: 14421-14426.

614 Schouten, S., Hopmans, E.C., Schefuß, E., and Sinninghe Damsté, J.S. (2002)
615 Distributional variations in marine crenarchaeotal membrane lipids: a new tool for
616 reconstructing ancient sea water temperatures? Earth Planet Sci Letters 204: 265-274.

617 Schouten, S., Forster, A., Panato, E., and Sinninghe Damsté, J. S. (2007) Towards the
618 calibration of the TEX86 paleothermometer in ancient green house worlds. Org
619 Geochem 38: 1537-1546.

620 Schouten, S., Hopmans, E.C., Baas, M., Boumann, H., Standfest, S., Könneke, M. et al
621 (2008a) Intact Membrane Lipids of "Candidatus Nitrosopumilus maritimus," a
622 Cultivated Representative of the Cosmopolitan Mesophilic Group I Crenarchaeota.
623 Appl Environ Microbiol 74: 2433-2440.

624 Schouten, S., van der Meer, M.T.J., Hopmans, E.C., and Sinninghe Damsté, J.S.
625 (2008b) Comment on "Lipids of marine Archaea: patterns and provenance in the water-
626 column and sediments" by Turich et al. (2007). Geochim Cosmochim Acta 72: 5432-
627 5436.

628 Schouten, S., Pitcher, A., Hopmans, E.C., Villanueva, L., van Bleijswijk, J., Sinninghe
629 Damsté, J.S. (2012) Intact polar and core glycerol dibiphytanyl glycerol tetraether lipids
630 in the Arabian Sea oxygen minimum zone: I. Selective preservation and degradation in
631 the water column and consequences for the TEX86. Geochim Cosmochim Acta 98:
632 228-243.

633 Schouten, S, Hopmans, E.C., and Sinninghe Damsté, J.S. (2013) The organic
634 geochemistry of glycerol dialkyl glycerol tetraether lipids: A review. *Org. Geochem.*
635 54: 19-61.

636 Shah, S.R., Mollenhauer, G., Ohkouchi, N., Eglinton, T.I., and Pearson, A. (2008)
637 Origins of archaeal tetraether lipids in sediments: Insights from radiocarbon analysis.
638 Geochim Cosmochim Acta 72: 4577-4594.

639 Sinninghe Damsté, J.S., Schouten, S., Hopmans, E.C., van Duin, A.C.T., and
640 Geenevasen, J.A.J. (2002) Crenarchaeol: the characteristic core glycerol dibiphytanyl
641 glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. J Lipid Res
642 43: 1641-1651.

643 Sinninghe Damsté, J.S., Rijpstra, W.I.C., Hopmans, E.C., Jung, M.-Y., Kim, J.-G.,
644 Rhee, S.-K. et al (2012) Intact Polar and Core Glycerol Dibiphytanyl Glycerol
645 Tetraether Lipids of Group I.1a and I.1b Thaumarchaeota in Soil. Appl Environ
646 Microbiol 78: 6866-6874.

647 Sintes, E., Bergauer, K., De Corte, D., Yokokawa, T., and Herndl, G.J. (2013) Archaeal
648 amoA gene diversity points to distinct biogeography of ammonia-oxidizing
649 Crenarchaeota in the ocean. Environ Microbiol 15: 1647-1658.

650 Spang, A., Hatzenpichler, R., Brochier-Armanet, C., Rattei, T., Tischler, P., Spieck, .E
651 et al (2010) Distinct gene set in two different lineages of ammonia-oxidizing archaea
652 supports the phylum Thaumarchaeota. Trends Microbiol 18: 331-340.

653 Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011)
654 MEGA5: molecular evolutionary genetics analysis using maximum likelihood,
655 evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-
656 2739.

657 Taylor, K.W.R., Huber, M., Hollis, C.J., Hernandez-Sanchez, M.T., and Pancost, R.D.
658 (2013) Re-evaluating modern and Paleogene GDGT distributions: Implications for SST
659 reconstructions. Global Planet Change 108: 158-174.

660 Turich, C., Freeman, K.H., Bruns, M.A., Conte, M., Jones, A.D., and Wakeham, S.G.
661 (2007) Lipids of marine Archaea: Patterns and provenance in the water-column and
662 sediments. Geochim Cosmochim Acta 71: 3272-3291.

663 Villanueva, L., Sinninghe Damsté, J.S., and Schouten S. A re-evaluation of the archaeal
664 membrane lipid biosynthetic pathway. Nat. Rev. Microbiol., in press.

665 Wakeham, S.G., Lewis, C.M., Hopmans. E.C., Schouten, S., and Sinninghe Damsté,
666 J.S. (2003) Archaea mediate anaerobic oxidation of methane in deep euxinic waters of
667 the Black Sea. Geochim Cosmichim Acta 67: 1359-1374.

668 Wang, S., Xiao, X., Jiang, L., Peng, X., Zhou, H., Meng, J. et al (2009) Diversity and
669 abundance of ammonia-oxidizing archaea in hydrothermal vent chimneys of the juan de
670 fuca ridge. Appl Environ Microbiol 75: 4216-4220.

671 Wuchter, C., Schouten, S., Coolen, M.J.L., and Sinninghe Damsté, J.S. (2004)
672 Temperature-dependent variation in the distribution of tetraether membrane lipids of
673 marine Crenarchaeota: Implications for TEX_{86} paleothermometry. Paleoceanog 19:
674 PA4028.

675 Wuchter, C., Schouten, S., Wakeham, S.G., and Sinninghe Damsté, J.S. (2005)
676 Temporal and spatial variation in tetraether membrane lipids of marine Crenarchaeota in
677 particulate organic matter: Implications for TEX_{86} paleothermometry. Paleoceanog 20:
678 PA3013.

679 Wuchter, C., Abbas, B., Coolen, M.J., Herfort, L., van Bleijswijk, J., Timmers, P. et al
680 (2006) Archaeal nitrification in the ocean. Proc Nat Acad Sci USA 103: 123171-2322.

Yakimov, M.M., Cono, V.L., Smedile, F., DeLuca, T.H., Juarez, S., Ciordia, S. et al (2011) Contribution of crenarchaeal autotrophic ammonia oxidizers to the dark primary production in Tyrrhenian deep waters (Central Mediterranean Sea). ISME J 5: 945-961.

Figure legends

Figure 1. Neighbor-joining tree of *amo*A protein sequences recovered from SPM at 170 m (in blue; 27 clones), SPM at 1050 m (in red; 27 clones), surface sediment (upper 0.25 cm) of station P3000 (3003 m water depth; in green; 32 clones) constructed with the Neighbor-Joining method (Saitou and Nei, 1987). Scale bar indicates 2% sequence dissimilarity. Clusters of Water column A/Sediments ('shallow water' marine clade) and B ('deep water" marine clade) of the *amo*A gene were defined by Francis *et al*. (2005). The evolutionary distances were computed using the Poisson correction method with a bootstrap test of 1,000 replicates (values higher than 50% are shown on the branches). The analysis involved 200 amino acid sequences and a total of 212 positions.

Figure 2. (A) Alignment of ammonia monooxygenase protein sequences, and (B) distribution of ammonia monooxygenase 'shallow' (blue) and 'deep (red) water' clusters in the ocean as determined from the analysis of metagenomic databases. In (A) 160 out of 216 amino acids (74% of total *amo*A protein) are shown. Amino acid residues defining the 'shallow' and 'deep water' *amo*A clusters are indicated with a star. Key: N, asparagine; S, serine; L, leucine; I, isoleucine; A, alanine; G, glycine; V, valine; R, arginine; K, lysine; F, phenylalanine; Q, glutamine; T, threonine; Y, tyrosine; H, histidine; C, cysteine; M, methionine. Amino acid residues shaded in the same color in the alignment are either identical or of the same nature. Pie charts in (B) indicate the distribution of the *amo*A 'shallow' (blue) and 'deep (red) water' clusters in metagenomic databases (*n* indicates number of sequences analyzed).

Figure 3. Phylogenetic tree of putative partial GGGP synthases recovered from the Arabian Sea OMZ SPM at seven different depths (varying from 20 to 2000 m), and surface sediment (upper 0.25 cm) of station P3000 (3003 water depth) revealing 'shallow' and 'deep water' clades. The analysis involved 334 amino acid sequences with 76 amino acid residues. The phylogenetic tree was inferred by maximum likelihood with the LG+G model of protein evolution. Branch support was calculated with the approximate likelihood ratio test (aLRT) and indicated on the branches. The scale bar indicates evolutionary distance of 0.1 substitutions per site.

Figure 4. (A) Alignment of partial GGGP synthases in different species of Thaumarchaeota, protein sequences from the Arabian Sea, and fosmids from the deep sea, and (B) distribution of GGGP synthase clusters in metagenomic databases with depth. In (A) the alignment of 76 amino acid residues between the 101–177 position of the GGGP synthase sequence of 'Ca. Nitrosoarchaeum koreensis'(ZP_08668912) is shown. See legend Figure 2 for details.

Figure 5. Depth profiles of (A) GDGT-2/GDGT-3 ratio of intact polar lipids in the Arabian Sea SPM ; (B) Percentage (%) of GGGP synthase sequences of the 'deep water' cluster detected in the Arabian Sea SPM; (C) GDGT-2/GDGT-3 ratio in marine surface sediments based on data reported by Kim *et al*., (2010); (D) Percentage (%) of GGGP synthase of the 'deep water' cluster in the marine metagenomes analyzed in this study (Table S2 and S3; Figure 4B).

Supporting information

Figure S1. Isoprenoid glycerol dialkyl glycerol tetraether (GDGT) structures.

Figure S2. Fractional abundance of GDGT-1, 2, 3, and crenarchaeol regioisomer in depth intervals in (A) Arabian Sea SPM, IPL-GDGT; (B) Arabian Sea SPM, CL-GDGT data; (C) Equatorial Pacific station 6, S1 (winter 1992); (D) Equatorial Pacific station 6, S2 (summer 1992) (Wuchter *et al*., 2005).

Figure S3. Abundance of archaeal and thaumarchaeotal 16S rRNA gene copies per liter in the Arabian Sea SPM as reported by Pitcher *et al*., 2011a and Schouten *et al*., 2012.

Figure S4. GDGT-2/GDGT-3 ratio in the Arabian Sea and Equatorial Pacific SPM. CL indicates GDGT core lipids derived from dead material and IPL indicates intact polar lipid-GDGTs derived from living biomass.

Table S1. Summary of physicochemical conditions and *amo*A gene quantification and gene expression in the Arabian Sea SPM and surface sediment reported in this study. Data from Pitcher *et al*., 2011a, Kraal *et al*., 2012, and this study.

Table S2. Metagenome projects from marine microbiomes in IMG/M database used in this study.

Table S3. Metagenome projects from Camera database included in the 'deep water' database.

Table S4. GGGP synthase gene primers designed and applied in this study.

Table S1. Summary of the physicochemical conditions, *amo*A gene quantification, and gene expression in the Arabian Sea SPM and surface sediment reported in this study.

Depth (m) SPM	T (°C)	NH_4^+ (μM)	NO_2^- (μM)	NO_3^-(μM)	HPO_4^{2-} (μM)	Oxygen (μM)	*amo*A gene copies L^{-1}	*amo*A mRNA L^{-1}
20	24.8	0.089	0.04	3.3	0.60	187.9	2.8E+07	3.1E+03
170	19.0	0.140	0.62	21.7	2.24	4.8	3.1E+08	5.5E+03
300	15.6	0.014	0.03	24.2	2.41	6.0	1.3E+08	1.2E+03
450	13.6	0.026	0.02	24.4	2.56	3.6	1.1E+08	2.2E+03
600	12.0	0.058	0.50	26.7	2.77	3.4	4.5E+07	1.3E+01
750	11.0	0.047	0.02	32	2.88	3.4	6.3E+07	0.0E+00
900	9.9	0.097	0.02	35.4	2.55	4.2	1.1E+08	9.4E+02
1050	8.5	0.098	0.02	37.9	3.08	4.6	2.1E+08	1.3E+03
1200	7.5	0.076	0.02	39.6	3.17	9.8	1.3E+08	1.1E+03
1350	6.5	0.066	0.02	39.3	3.19	18.8	6.6E+07	7.1E+02
1500	5.5	0.004	0.01	39.7	3.19	28.5	8.4E+07	1.0E+03
2000	3.2	0.043	0.03	38.1	2.98	62.7	6.0E+07	2.1E+02
Surface sediment	**T (°C) ****	**NH_4^+ (μM)**	**NO_2^- (μM)**	**NO_3^-(μM)**	**HPO_4^{2-} (μM)**	**Oxygen (μM)**	***amo*A gene* copies g dw^{-1}**	***amo*A* mRNA g dw^{-1}**
3003 m (P3000)	1.4	55.6	8.3	46.2	3.8	70	3.9E+8 (2.5E+7)	1.8E+4 (3E+3)

Data by Pitcher *et al*., 2011a and Kraal *et al*., 2012; Biogeosciences 9:2603 (pore water nutrients in the upper 0.5 cm). *Data obtained in this study. Dw = dry weight. †nd = not detected. Numbers between parentheses are standard deviations of three QPCR measurements. Limit of detection of *amo*A gene copies in the surface sediment = 200 gene copies g^{-1} dry weight. **values in bottom water due to lack of measurement. Quantitative PCR of the *amo*A gene in surface sediments was performed as described by Pitcher *et al*. (2011a).

Table S2. Metagenome projects from Marine microbiomes in IMG/M used in this study.

Taxon_id	Proposal Name	Genome Name / Sample Name
2040502005	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, Sample 10335 (Summer fosmids)
3300000149	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2009 P16 10m (Line P August 2009 P16 10m, March 2012 Assem)
3300000171	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 47 07/07/10 200m (Saanich Inlet 47 07/07/10 200m, March 2012 Assem)
3300000146	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 54 02/08/11 120m (Saanich Inlet 54 02/08/11 120m, March 2012 Assem)
2014613002	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	1_Upper_euphotic
3300000186	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2010 P16 2000m (Line P February 2010 P16 2000m, April 2012 Assem)
3300000218	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P20 2000m (Line P June 2009 P20 2000m, March 2012 Assem)
3300000198	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_200 (Line P sample_F_10_SI03_200, March 2012 Assem)
3300000216	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 150m (Saanich Inlet 53 01/11/11 150m, March 2012 Assem)
3300000174	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 60 08/10/11 200m (Saanich Inlet 60 08/10/11 200m, April 2012 Assem)

2189573014	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_09_P20_1000 (sample_J_09_P20_1000 June 2011 assem)
3300000147	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 54 02/08/11 150m (Saanich Inlet 54 02/08/11 150m, April 2012 Assem)
3300000141	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P4 1300m (Line P June 2008 P4 1300m, March 2012 Assem)
3300000256	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_120 (Line P sample_F_10_SI03_120, March 2012 Assem)
3300000222	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P12 500m (Line P June 2009 P12 500m, March 2012 Assem)
2061766003	Guaymas Basin hydrothermal plume	Hydrothermal vent microbial communities from Guaymas and Carmen Basins, Gulf of California, Sample 457
3300000167	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 120m (Saanich Inlet 39 11/10/09 120m, March 2012 Assem)
2077657013	Marine Bacterioplankton communities from the Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Summer (Summer fosmids Sept 2010 assemblies)
3300000158	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 54 02/08/11 100m (Saanich Inlet 54 02/08/11 100m, March 2012 Assem)
3300000266	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_09_P20_500 (Line P sample_J_09_P20_500, March 2012 Assem)
2149837028	Deepwater Horizon Subsurface Plume Metatranscriptome	52-4 In plume (52-4 In Plume)
3300000250	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2009 P26 1000m (Line P February 2009 P26 1000m, March 2012 Assem)

3300000134	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBA sample SWE 07_21m (Baltic Sea site KBA sample SWE 07_21m, Oct 2011 Assem)
3300000214	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 54 02/08/11 200m (Saanich Inlet 54 02/08/11 200m, March 2012 Assem)
2189573017	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_120 (sample_F_10_SI03_120 June 2011 assem)
3300000118	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site OR sample ARG 06_12.3m (Tierra del Fuego site OR sample ARG 06_12.3m, Oct 2011 Assem)
3300000265	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_10 (Line P sample_A_09_P04_10, April 2012 Assem)
3300000195	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 150m (Saanich Inlet 39 11/10/09 150m, March 2012 Assem)
2014642000	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	6_Upper_euphotic
3300000131	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 02_11.3m (Tierra del Fuego site MC sample ARG 02_11.3m, Jan 2012 Assem)
2189573010	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P20_1000 (sample_A_09_P20_1000 June 2011 assem)
2014613003	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	2_Base_of_chrolophyll_max
3300000168	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P12 10m (Line P June 2009 P12 10m, March 2012 Assem)
3300000133	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 02_45m (Svalbard Archipelago station 1 sample NOR 02_45m, Jan 2012 Assem)
3300000116	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March

		2010 (Delaware MO Spring March 2010, Nov 2011 assem)
2189573015	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample F_10_SI03_10 (sample_F_10_SI03_10 June 2011 assem)
2014642003	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	7_Oxygen_minimum_layer
2156126013	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P20_1000 (A_09_P20_1000)
3300000159	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P26 10m (Line P August 2008 P26 10m, March 2012 Assem)
2149837027	Deepwater Horizon Subsurface Plume Metatranscriptome	52-1 Below Plume (52-1 Below Plume)
3300000144	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P4 1000m (Line P June 2008 P4 1000m, March 2012 Assem)
3300000172	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 200m (Saanich Inlet 34 06/16/09 200m, May 2012 Assem)
3300000153	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 135m (Saanich Inlet 39 11/10/09 135m, April 2012 Assem)
3300000192	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 60 08/10/11 100m (Saanich Inlet 60 08/10/11 100m, March 2012 Assem)
3300000136	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 02_9.5m (King George Island site S1 sample ANT 02_9.5m, Dec 2011 Assem)
3300000237	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 150m (Saanich Inlet 34 06/16/09 150m, March 2012 Assem)
3300000127	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 05_45m (Svalbard Archipelago station 1 sample NOR 05_45m, Nov 2011 Assem)

3300000322	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P12 1000m (Line P August 2008 P12 1000m, June 2012 Assem)
3300000224	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 10m (Saanich Inlet 34 06/16/09 10m, March 2012 Assem)
2189573012	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_08_P26_500 (sample_J_08_P26_500 June 2011 assem)
2189573007	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_1000 (A_09_P04_1000 June 2011 assem)
3300000161	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P12 2000m (Line P August 2008 P12 2000m, March 2012 Assem)
3300000261	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P20_1000 (Line P sample_A_09_P20_1000, April 2012 Assem)
2008193000	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, sample from Summer (Summer fosmid end sequences)
3300000239	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 36 08/11/09 120m (Saanich Inlet 36 08/11/09 120m, March 2012 Assem)
2162886005	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_135 (F_10_SI03_135)
2189573009	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_10 (sample_A_09_P04_10 June 2011 assem)
3300000163	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P16 2000m (Line P June 2009 P16 2000m, March 2012 Assem)
3300000254	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 100m (Saanich Inlet 34 06/16/09 100m, March 2012 Assem)

3300000226	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 135m (Saanich Inlet 34 06/16/09 135m, March 2012 Assem)
3300000123	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 06_23.45m (King George Island site S2 sample ANT 06_23.45m, Oct 2011 Assem)
3300000199	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 10m (Saanich Inlet 39 11/10/09 10m, March 2012 Assem)
2156126011	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_500 (A_09_P04_500)
2156126009	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample F_10_SI03_10 (F_10_S103_10)
3300000242	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site OR sample ARG 05_12.3m (Tierra del Fuego site OR sample ARG 05_12.3m, Oct 2011 Assem)
3300000121	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 03_11.3m (Tierra del Fuego site MC sample ARG 03_11.3m, Oct 2011 Assem)
3300000324	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 100m (Saanich Inlet 48 08/11/10 100m, June 2012 Assem)
3300000248	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2009 P12 500m (Line P February 2009 P12 500m, March 2012 Assem)
3300000170	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 36 08/11/09 135m (Saanich Inlet 36 08/11/09 135m, March 2012 Assem)
3300000183	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P20 500m (Line P August 2008 P20 500m, March 2012 Assem)
3300000155	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 36 08/11/09 200m (Saanich Inlet 36 08/11/09 200m, March 2012 Assem)

2162886004	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_100 (F_10_SI03_100)
3300000263	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_1000 (Line P sample_A_09_P04_1000, March 2012 Assem)
3300000164	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 200m (Saanich Inlet 39 11/10/09 200m, May 2012 Assem)
3300000151	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 200m (Saanich Inlet 53 01/11/11 200m, March 2012 Assem)
3300000249	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2009 P12 1000m (Line P February 2009 P12 1000m, March 2012 Assem)
3300000125	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 01_11.3m (Tierra del Fuego site MC sample ARG 01_11.3m, Nov 2011 Assem)
3300000143	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 10m (Saanich Inlet 53 01/11/11 10m, March 2012 Assem)
3300000128	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 08_45m (Svalbard Archipelago station 1 sample NOR 08_45m, Dec 2011 Assem)
3300000129	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 04_23.45m (King George Island site S2 sample ANT 04_23.45m, Dec 2011 Assem)
2156126010	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_1300 (A_09_P04_1300)
3300000140	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2009 P26 500m (Line P February 2009 P26 500m, March 2012 Assem)
3300000188	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 60 08/10/11 150m (Saanich Inlet 60 08/10/11 150m, March 2012 Assem)

2040502004	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, Sample 10334 (Winter fosmids)
2008193001	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, sample from Winter (Winter fosmid end sequences)
2156126012	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_10 (A_09_P04_10)
2140918005	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from high methane PC12-225-485cm (High methane PC12-225-485cm Jan 2011 assembly)
2189573013	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_09_P20_500 (sample_J_09_P20_500 June 2011 assem)
3300000264	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_500 (Line P sample_A_09_P04_500, March 2012 Assem)
3300000247	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2009 P26 500m (Line P August 2009 P26 500m, March 2012 Assem)
3300000166	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 200m (Saanich Inlet 48 08/11/10 200m, April 2012 Assem)
3300000213	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_150 (Line P sample_F_10_SI03_150, April 2012 Assem)
3300000157	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P26 1000m (Line P August 2008 P26 1000m, March 2012 Assem)
3300000323	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2009 P20 2000m (Line P August 2009 P20 2000m, June 2012 Assem)
2199352009	Marine subseafloor sediment microbial communities from Peru Margin, Ocean Drilling Program Site 1229	Marine subseafloor sediment microbial communities, sample from White Oak River

		Estuary, NC, USA 14E (White Oak River Estuary June 2011 assem)
3300000130	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 15_50m (Svalbard Archipelago station 2 sample NOR 15_50m, Dec 2011 Assem)
3300000219	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2010 P16 1000m (Line P February 2010 P16 1000m, May 2012 Assem)
3300000252	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P16 1000m (Line P June 2008 P16 1000m, March 2012 Assem)
3300000135	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 03_9.5m (King George Island site S1 sample ANT 03_9.5m, Dec 2011 Assem)
2189573018	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_135 (sample_F_10_SI03_135 June 2011 assem)
3300000119	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 01_9.5m (King George Island site S1 sample ANT 01_9.5m, Oct 2011 Assem)
3300000207	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 10m (Saanich Inlet 48 08/11/10 10m, March 2012 Assem)
3300000255	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_135 (Line P sample_F_10_SI03_135, March 2012 Assem)
3300000187	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 100m (Saanich Inlet 53 01/11/11 100m, March 2012 Assem)
3300000115	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Summer July 2011 (Delaware MO Summer July 2011, Nov 2011 assem)
3300000152	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P12 500m (Line P June 2008 P12 500m, May 2012 Assem)

3300000211	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 135m (Saanich Inlet 53 01/11/11 135m, March 2012 Assem)
3300000262	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_1300 (Line P sample_A_09_P04_1300, March 2012 Assem)
3300000259	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_08_P26_500 (Line P sample_J_08_P26_500, March 2012 Assem)
3300000148	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 47 07/07/10 100m (Saanich Inlet 47 07/07/10 100m, March 2012 Assem)
3300000215	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 53 01/11/11 120m (Saanich Inlet 53 01/11/11 120m, March 2012 Assem)
3300000204	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 36 08/11/09 150m (Saanich Inlet 36 08/11/09 150m, March 2012 Assem)
2189573019	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_150 (sample_F_10_SI03_150 June 2011 assem)
3300000132	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 05_23.45m (King George Island site S2 sample ANT 05_23.45m, Jan 2012 Assem)
3300000258	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_09_P20_1000 (Line P sample_J_09_P20_1000, April 2012 Assem)
3300000200	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 150m (Saanich Inlet 48 08/11/10 150m, March 2012 Assem)
3300000209	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2008 P20 2000m (Line P August 2008 P20 2000m, March 2012 Assem)
3300000225	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 34 06/16/09 120m (Saanich Inlet 34 06/16/09 120m, March 2012 Assem)

3300000251	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P16 500m (Line P June 2008 P16 500m, March 2012 Assem)
2189573006	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_500 (sample_A_09_P04_500 June 2011 assem)
3300000173	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P February 2010 P16 500m (Line P February 2010 P16 500m, March 2012 Assem)
3300000193	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 47 07/07/10 135m (Saanich Inlet 47 07/07/10 135m, March 2012 Assem)
2189573016	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_100 (sample_F_10_SI03_100 June 2011 assem)
3300000196	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2009 P16 2000m (Line P August 2009 P16 2000m, March 2012 Assem)
2014642002	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	5_Below_upper_mesopelagic
3300000260	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P20_500 (Line P sample_A_09_P20_500, March 2012 Assem)
3300000241	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBB sample SWE 21_20.5m (Baltic Sea site KBB sample SWE 21_20.5m, Oct 2011 Assem)
2014642004	Marine planktonic communities from Hawaii Ocean Times Series Station (HOT/ALOHA)	4_Deep_abyss
2149837026	Deepwater Horizon Subsurface Plume Metatranscriptome	16-5 In Plume (16-5 In Plume)
2264265093	Marine Bacterioplankton communities from the Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Summer
3300000212	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 47 07/07/10 120m (Saanich Inlet 47 07/07/10 120m, March 2012 Assem)

3300000181	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P4 500m (Line P June 2008 P4 500m, March 2012 Assem)
3300000126	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBB sample SWE 26_20.5m (Baltic Sea site KBB sample SWE 26_20.5m, Nov 2011 Assem)
2189573011	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P20_500 (sample_A_09_P20_500 June 2011 assem)
3300000150	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 120m (Saanich Inlet 48 08/11/10 120m, March 2012 Assem)
3300000190	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P16 1000m (Line P June 2009 P16 1000m, March 2012 Assem)
3300000179	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P16 500m (Line P June 2009 P16 500m, March 2012 Assem)
3300000221	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P12 2000m (Line P June 2008 P12 2000m, March 2012 Assem)
3300000243	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 18_50m (Svalbard Archipelago station 2 sample NOR 18_50m, Dec 2011 Assem)
3300000325	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 39 11/10/09 100m (Saanich Inlet 39 11/10/09 100m, June 2012 Assem)
3300000120	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 13_50m (Svalbard Archipelago station 2 sample NOR 13_50m, Oct 2011 Assem)
2162886003	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_J_08_P26_500 (J_08_P26_500)
3300000154	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 47 07/07/10 150m (Saanich Inlet 47 07/07/10 150m, March 2012 Assem)

3300000201	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 54 02/08/11 135m (Saanich Inlet 54 02/08/11 135m, March 2012 Assem)
3300000238	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 36 08/11/09 100m (Saanich Inlet 36 08/11/09 100m, March 2012 Assem)
3300000137	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample F_10_SI03_10 (Line P sample_F_10_SI03_10, March 2012 Assem)
3300000160	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Saanich Inlet 48 08/11/10 135m (Saanich Inlet 48 08/11/10 135m, March 2012 Assem)
3300000117	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Winter December 2010 (Delaware MO Winter December 2010, Nov 2011 assem)
3300000124	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBA sample SWE 12_21m (Baltic Sea site KBA sample SWE 12_21m, Oct 2011 Assem)
3300000253	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2008 P12 1000m (Line P June 2008 P12 1000m, April 2012 Assem)
3300000101	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Early Summer May 2010 (Delaware MO Early Summer May 2010, Feb 2012 assem)
2166559025	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_A_09_P04_1300 (A_09_P04_1300 June 2011 assembly)
2189573008	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P sample_F_10_SI03_200 (sample_F_10_SI03_200 June 2011 assem)
3300000142	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P August 2009 P16 500m (Line P August 2009 P16 500m, March 2012 Assem)

2077657020	Marine Bacterioplankton communities from the Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Winter (Winter fosmids Sept 2010 assemblies)
3300000223	Marine microbial communities from the Eastern Subtropical North Pacific Ocean, Expanding Oxygen minimum zones	Line P June 2009 P4 10m (Line P June 2009 P4 10m, March 2012 Assem)

Table S3. Metagenome projects from Camera included in the 'deep water' custom database.

Accession number	Project	Depth (m)	Name	Description
CAM_SMPL_000813	Moore Marine Phage/Virus Metagenomes	1000	1000 meters DNA	
CAM_SMPL_000829	Moore Marine Phage/Virus Metagenomes	1000	1000 meters RNA	
CAM_SMPL_000807	Moore Marine Phage/Virus Metagenomes	805	12C Fraction ANME Virome	Santa Monica Basin, offshore Los Angeles, CA
ALVINELLA_SMPL_20041130	Alvinella pompejana Epibiont Metagenome	2500	ALVINELLA	ALVINELLA - Alvinella Pompejana Epibionts
BATS_SMPL_174-1	Metagenomic Analysis of the North Atlantic Spring Bloom	4857	BATS_SMPL_174-1	BATS-174-1
BATS_SMPL_174-2	Metagenomic Analysis of the North Atlantic Spring Bloom	4857	BATS_SMPL_174-2	BATS-174-2
BATS_SMPL_179-1	Metagenomic Analysis of the North Atlantic Spring Bloom	2687	BATS_SMPL_179-1	BATS-179-1
BATS_SMPL_179-2	Metagenomic Analysis of the North Atlantic Spring Bloom	2687	BATS_SMPL_179-2	BATS-179-2
CAM_SMPL_001118	Microbial Oceanography Course	500	BDAAmpliconsMO2009.MO-09-1_C7-N18	
CAM_SMPL_001119	Microbial Oceanography Course	800	BDAAmpliconsMO2009.MO-09-1_C7-N19	

CAM_SMPL_001128	Microbial Oceanography Course	1000	BDAAmpliconsMO2009.MO-09-3_C2-N24	
CAM_SMPL_001138	Microbial Oceanography Course	500	BDAAmpliconsMO2009.MO10-1C1N24	
CAM_SMPL_000799	Moore Marine Phage/Virus Metagenomes	1970	Black Sea Sediment Metagenome	
DEEPMED_SMPL_KM3_20041117	Mediterranean Bathypelagic Habitat Metagenome	3010	DEEPMED	DEEPMED - Mediterranean Bathypelagic Habitat
DEEPMED_SMPL_KM3_20041117	Mediterranean Bathypelagic Habitat Metagenome	3010	DEEPMED	DEEPMED - Mediterranean Bathypelagic Habitat
CAM_S_466	Guaymas Basin deep-sea Metagenome	2040	GD1	
CAM_S_465	Guaymas Basin deep-sea Metagenome	2771	GD10	
CAM_S_469	Guaymas Basin deep-sea Metagenome	2040	GD2	
CAM_S_468	Guaymas Basin deep-sea Metagenome	2771	GD5	
CAM_S_463	Guaymas Basin deep-sea Metagenome	2050	GD6	
CAM_S_464	Guaymas Basin deep-sea Metagenome	2060	GD7	
CAM_S_467	Guaymas Basin deep-sea Metagenome	2771	GD8	
CAM_S_462	Guaymas Basin deep-sea Metagenome	2060	GD9	
JGI_SMPL_HF4000_12-21-03	Microbial Community Genomics at the HOT/ALOHA	4000	HF4000_12-21-03	HF4000_12-21-03 - HOT station ALOHA, 4000 m
JGI_SMPL_HF500_10-06-02	Microbial Community Genomics at the HOT/ALOHA	500	HF500_10-06-02	HF500_10-06-02 - HOT station ALOHA, 500 m

JGI_SMPL_HF770_12-21-03	Microbial Community Genomics at the HOT/ALOHA	770	HF770_12-21-03	HF770_12-21-03 - HOT station ALOHA, 770 m
HF_SMPL_HOT179_500M_CDNA	Microbial Community Genomics at the HOT/ALOHA	500	HF_SMPL_HOT179_500M_CDNA	
HF_SMPL_HOT179_500M_GDNA	Microbial Community Genomics at the HOT/ALOHA	500	HF_SMPL_HOT179_500M_GDNA	
HF_SMPL_HOT179_500M_SG	Microbial Community Genomics at the HOT/ALOHA	500	HF_SMPL_HOT179_500M_SG	HOT179_500m_Shotgun
HF_SMPL_HOT186_500M_GDNA	Microbial Community Genomics at the HOT/ALOHA	500	HF_SMPL_HOT186_500M_GDNA	HOT186_500m_gDNA
MF_SMPL_ABOONEI	Moore Marine Microbial Sequencing	1800	MF_ABOONEI	MF_ABOONEI - Aciduliprofundum boonei T469
MF_SMPL_BOGUAY	Moore Marine Microbial Sequencing	2000	MF_BOGUAY	MF_BOGUAY - Beggiatoa sp. 'Orange Guaymas'
MF_SMPL_CAT7	Moore Marine Microbial Sequencing	2500	MF_CAT7	MF_CAT7 - Carnobacterium sp. AT7
MF_SMPL_CDSM653	Moore Marine Microbial Sequencing	1395	MF_CDSM653	MF_CDSM653 - Caldanaerobacter DSM 12653
MF_SMPL_CMTB2	Moore Marine Microbial Sequencing	2300	MF_CMTB2	MF_CMTB2 - Caminibacter mediatlanticus TB-2
MF_SMPL_DSM11836	Moore Marine Microbial Sequencing	3550	MF_DSM11836	MF_DSM11836 - Thermococcus MP DSMZ 11836
MF_SMPL_DSM6158	Moore Marine Microbial Sequencing	2000	MF_DSM6158	MF_DSM 6158 - Pyrodictium abyssi DSM 6158
MF_SMPL_HG1285	Moore Marine Microbial Sequencing	2200	MF_HG1285	MF_HG1285 - Hydrogenivirga sp. 128-5-R1-1
MF_SMPL_MADE	Moore Marine Microbial Sequencing	1000	MF_MADE	MF_MADE - Alteromonas macleodii 'Deep ecotype'
MF_SMPL_MPKA3	Moore Marine Microbial Sequencing		MF_MPKA3	MF_MPKA3 - Marinitoga piezophila KA3

MF_SMPL_OBOE	Moore Marine Microbial Sequencing	2895	_OBOE	MF_OBOE - Oceanospirillales bacterium 'Osedax endosymbiont'
MF_SMPL_PE36	Moore Marine Microbial Sequencing	3600	MF_PE36	MF_PE36 - Moritella sp. PE36
MF_SMPL_SPV1	Moore Marine Microbial Sequencing	1200	MF_SPV1	MF_SPV1 - Mariprofundus ferrooxydans PV-1
MF_SMPL_TAM4	Moore Marine Microbial Sequencing	2600	MF_TAM4	MF_TAM4 - Thermococcus sp. AM4
CAM_SMPL_SRA022158	Antarctica Aquatic Microbial Metagenome	1320	Station_354 -- 0.1 um	0.1um sequencing
CAM_SMPL_SRA022159	Antarctica Aquatic Microbial Metagenome	1320	Station_354 -- 0.8 um	0.8um sequencing
CAM_SMPL_SRA022093	Antarctica Aquatic Microbial Metagenome	1320	Station_354 -- 3 um	3.0um sequencing
CAM_SMPL_SRA022161	Antarctica Aquatic Microbial Metagenome	890	Station_356 -- 0.1 um	0.1um sequencing
CAM_SMPL_SRA022162	Antarctica Aquatic Microbial Metagenome	890	Station_356 -- 0.8 um	0.8um sequencing
CAM_SMPL_SRA022096	Antarctica Aquatic Microbial Metagenome	890	Station_356 -- 3 um	3.0um sequencing
CAM_SMPL_SRA022170	Antarctica Aquatic Microbial Metagenome	1170	Station_361 -- 0.1 um	0.1um sequencing
CAM_SMPL_SRA022103	Antarctica Aquatic Microbial Metagenome	1170	Station_361 -- 0.8 um	MID barcoded 0.8um and 3.0um filters from the same station
CAM_SMPL_SRA022102	Antarctica Aquatic Microbial Metagenome	1170	Station_361 -- 3.0 um	MID barcoded 0.8um and 3.0um filters from the same station
CAM_SMPL_SRA022176	Antarctica Aquatic Microbial Metagenome	3693	Station_365 -- 0.1 um	0.1um sequencing
CAM_SMPL_SRA022109	Antarctica Aquatic Microbial Metagenome	3693	Station_365 -- 0.8 um	MID barcoded 0.8um and 3.0um filters from the same station
CAM_SMPL_SRA022108	Antarctica Aquatic Microbial Metagenome	3693	Station_365 -- 3.0 um	MID barcoded 0.8um and 3.0um filters from the same station
CAM_SMPL_000962	Moore Marine Phage/Virus Metagenomes	500	Uncultured virus Virus 3. San Pedro Ocean Time Series Microbial Observatory	

CAM_SMPL_000962	Moore Marine Phage/Virus Metagenomes	500	Uncultured virus Virus 3. San Pedro Ocean Time Series Microbial Observatory	
CAM_SMPL_000962	Moore Marine Phage/Virus Metagenomes	500	Uncultured virus Virus 3. San Pedro Ocean Time Series Microbial Observatory	
CAM_SMPL_001014	Moore Marine Phage/Virus Metagenomes	885	Uncultured virus Virus 4. San Pedro Ocean Time Series Microbial Observatory	
CAM_SMPL_001014	Moore Marine Phage/Virus Metagenomes	885	Uncultured virus Virus 4. San Pedro Ocean Time Series Microbial Observatory	
CAM_SMPL_001014	Moore Marine Phage/Virus Metagenomes	885	Uncultured virus Virus 4. San Pedro Ocean Time Series Microbial Observatory	
CAM_SMPL_001005	Moore Marine Phage/Virus Metagenomes	860	Uncultured virus Virus 5.	
CAM_SMPL_001005	Moore Marine Phage/Virus Metagenomes	860	Uncultured virus Virus 5.	
CAM_SMPL_000842	Moore Marine Phage/Virus Metagenomes	3670	VAGALB1/1	
CAM_SMPL_000839	Moore Marine Phage/Virus Metagenomes	805	Virome 13C-enriched ANME virome	Santa Monica Basin, offshore Los Angeles, CA
CAM_SMPL_000818	Moore Marine Phage/Virus Metagenomes	805	Virome ANME virome	Santa Monica Basin, offshore Los Angeles, CA
CAM_SMPL_000840	Moore Marine Phage/Virus Metagenomes	517	Virome Archaeal dominated cold seeps Costa Rica dsDNA virome	http://4dgeo.whoi.edu/webdata/DAQ/AT15-44/Alvin-D4510/Src1/Images0001/SubSea1.20090304_165800.jpg
CAM_SMPL_000805	Moore Marine Phage/Virus Metagenomes	517	Virome Archaeal dominated cold seeps Costa Rica RNA virome	http://4dgeo.whoi.edu/webdata/DAQ/AT15-44/Alvin-D4510/Src1/Images0001/SubSea1.20090304_165800.jpg
CAM_SMPL_000841	Moore Marine Phage/Virus Metagenomes	517	Virome Archaeal dominated cold seeps Costa Rica ssDNA virome	http://4dgeo.whoi.edu/webdata/DAQ/AT15-44/Alvin-

				D4510/Src1/Images0001/SubSea1.20090304_165800.jpg
CAM_SMPL_000845	Moore Marine Phage/Virus Metagenomes	2400	Virome BADE1	
CAM_SMPL_000718	Moore Marine Phage/Virus Metagenomes	2505	Virome EPR hydrothermal vent: Extracellular RNA virome	
CAM_SMPL_000719	Moore Marine Phage/Virus Metagenomes	2505	Virome EPR hydrothermal vent: Extracellular ssDNA virome	
CAM_SMPL_000720	Moore Marine Phage/Virus Metagenomes	2511	Virome EPR hydrothermal vent: Induced RNA virome	
CAM_SMPL_000721	Moore Marine Phage/Virus Metagenomes	2511	Virome EPR hydrothermal vent: Induced ssDNA virome	
CAM_SMPL_000834	Moore Marine Phage/Virus Metagenomes	1987	Virome Guaymas hydrothermal vent: Extracellular RNA virome	Placed near a Riftia patch near top, scale worms nearby, shimmering water. Rebecca's Roost
CAM_SMPL_000804	Moore Marine Phage/Virus Metagenomes	1987	Virome Guaymas hydrothermal vent: Extracellular ssDNA virome	Placed near a Riftia patch near top, scale worms nearby, shimmering water. Rebecca's Roost
CAM_SMPL_000809	Moore Marine Phage/Virus Metagenomes	1987	Virome Guaymas hydrothermal vent: Induced RNA virome	Placed near a Riftia patch near top, scale worms nearby, shimmering water. Rebecca's Roost
CAM_SMPL_000822	Moore Marine Phage/Virus Metagenomes	1987	Virome Guaymas hydrothermal vent: Induced ssDNA virome	Placed near a Riftia patch near top, scale worms nearby, shimmering water. Rebecca's Roost
CAM_SMPL_000832	Moore Marine Phage/Virus Metagenomes	580	Virome Methanogenic sediments virome	Santa Barbara Basin
CAM_SMPL_000979	Moore Marine Phage/Virus Metagenomes	1300	Virome Subarctic Pacific-10	
CAM_SMPL_000985	Moore Marine Phage/Virus Metagenomes	500	Virome Subarctic Pacific-2	
CAM_SMPL_000998	Moore Marine Phage/Virus Metagenomes	1000	Virome Subarctic Pacific-3	
CAM_SMPL_000956	Moore Marine Phage/Virus Metagenomes	2000	Virome Subarctic Pacific-4	

CAM_SMPL_000957	Moore Marine Phage/Virus Metagenomes	500	Virome Subarctic Pacific-5	
CAM_SMPL_001003	Moore Marine Phage/Virus Metagenomes	1000	Virome Subarctic Pacific-6	
CAM_SMPL_000986	Moore Marine Phage/Virus Metagenomes	2000	Virome Subarctic Pacific-7	
CAM_SMPL_001012	Moore Marine Phage/Virus Metagenomes	500	Virome Subarctic Pacific-8	
CAM_SMPL_000970	Moore Marine Phage/Virus Metagenomes	1000	Virome Subarctic Pacific-9	
CAM_SMPL_000833	Moore Marine Phage/Virus Metagenomes	805	Virome Suboxic marine basin virome	Santa Monica Basin, offshore Los Angeles, CA
CAM_SMPL_000843	Moore Marine Phage/Virus Metagenomes	3077	Virome Whittard Canyon (VAWC1/1)	
CAM_SMPL_WHALEFALLBONE	Whale Fall Metagenome	560	WHALEFALLBONE	WHALEFALLBONE - Whale fall carcass bone, W. Antarctic Peninsula Shelf
CAM_SMPL_WHALEFALLMAT	Whale Fall Metagenome	1674	WHALEFALLMAT	WHALEFALLMAT - Whale fall Santa Cruz Basin, USA

Table S4. GGGP synthase gene primers designed and applied in this study.

Primer name	Sequence 5'-3'
GGGP_Thaum_301F_short	ATGAAYTCDGARAAYCCNTA
GGGP_Thaum_301F_longdeg	ATGAAYTCDGARAAYCCNTAYT
GGGP_Thaum_301F_long	ATGAAYTCRGARAAYCCNTAYT
GGGP_Thaum_530R_long	GCYTCHARATANACRAAYCTCAT
GGGP_Thaum_530R_deg	GCYTCHARRTANACRAAYCKCAT
GGGP_Thaum_530R_short	GCYTCHARATANACRAAYCTCA
GGGP_Thaum_530R_ldeg	GCYTCHARATANACRAAYCTCAT

*Degenerated bases: Y (C/T); R (A/G); D (A/G/T); N (A/T/C/G); H (A/C/T). The only primer pair that gave a positive PCR amplification was GGGP_Thaum_301F_short/530R_short, and the optimum melting temperature was 48°C.

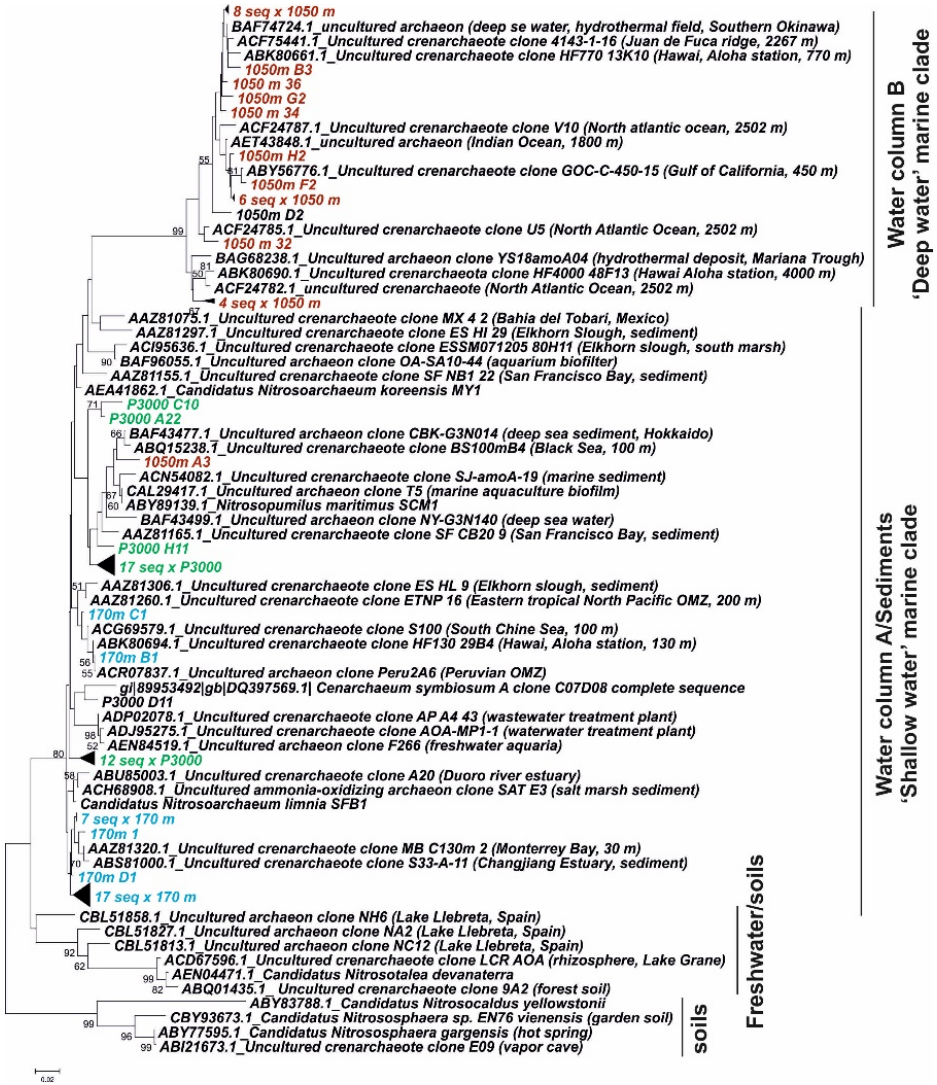

8 seq x 1050 m
BAF74724.1_uncultured archaeon (deep se water, hydrothermal field, Southern Okinawa)
ACF75441.1_Uncultured crenarchaeote clone 4143-1-16 (Juan de Fuca ridge, 2267 m)
ABK80661.1_Uncultured crenarchaeote clone HF770 13K10 (Hawai, Aloha station, 770 m)
1050m B3
1050 m 36
1050m G2
1050 m 34
ACF24787.1_Uncultured crenarchaeote clone V10 (North atlantic ocean, 2502 m)
AET43848.1_uncultured archaeon (Indian Ocean, 1800 m)
1050m H2
ABY56776.1_Uncultured crenarchaeote clone GOC-C-450-15 (Gulf of California, 450 m)
1050m F2
6 seq x 1050 m
1050m D2
ACF24785.1_Uncultured crenarchaeote clone U5 (North Atlantic Ocean, 2502 m)
1050 m 32
BAG68238.1_Uncultured archaeon clone YS18amoA04 (hydrothermal deposit, Mariana Trough)
ABK80690.1_Uncultured crenarchaeota clone HF4000 48F13 (Hawai Aloha station, 4000 m)
ACF24782.1_uncultured crenarchaeote (North Atlantic Ocean, 2502 m)
4 seq x 1050 m
AAZ81075.1_Uncultured crenarchaeote clone MX 4 2 (Bahia del Tobari, Mexico)
AAZ81297.1_Uncultured crenarchaeote clone ES HI 29 (Elkhorn Slough, sediment)
ACI95636.1_Uncultured crenarchaeote clone ESSM071205 80H11 (Elkhorn slough, south marsh)
BAF96055.1_Uncultured archaeon clone OA-SA10-44 (aquarium biofilter)
AAZ81155.1_Uncultured crenarchaeote clone SF NB1 22 (San Francisco Bay, sediment)
AEA41862.1_Candidatus Nitrosoarchaeum koreensis MY1
P3000 C10
P3000 A22
BAF43477.1_Uncultured archaeon clone CBK-G3N014 (deep sea sediment, Hokkaido)
ABQ15238.1_Uncultured crenarchaeote clone BS100mB4 (Black Sea, 100 m)
1050m A3
ACN54082.1_Uncultured crenarchaeote clone SJ-amoA-19 (marine sediment)
CAL29417.1_Uncultured archaeon clone T5 (marine aquaculture biofilm)
ABY89139.1_Nitrosopumilus maritimus SCM1
BAF43499.1_Uncultured archaeon clone NY-G3N140 (deep sea water)
AAZ81165.1_Uncultured crenarchaeote clone SF CB20 9 (San Francisco Bay, sediment)
P3000 H11
17 seq x P3000
AAZ81306.1_Uncultured crenarchaeote clone ES HL 9 (Elkhorn slough, sediment)
AAZ81260.1_Uncultured crenarchaeote clone ETNP 16 (Eastern tropical North Pacific OMZ, 200 m)
170m C1
ACG69579.1_Uncultured crenarchaeote clone S100 (South Chine Sea, 100 m)
ABK80694.1_Uncultured crenarchaeote clone HF130 29B4 (Hawai, Aloha station, 130 m)
170m B1
ACR07837.1_Uncultured archaeon clone Peru2A6 (Peruvian OMZ)
gi|89953492|gb|DQ397569.1| Cenarchaeum symbiosum A clone C07D08 complete sequence
P3000 D11
ADP02078.1_Uncultured crenarchaeote clone AP A4 43 (wastewater treatment plant)
ADJ95275.1_Uncultured crenarchaeote clone AOA-MP1-1 (waterwater treatment plant)
AEN84519.1_Uncultured archaeon clone F266 (freshwater aquaria)
12 seq x P3000
ABU85003.1_Uncultured crenarchaeote clone A20 (Duoro river estuary)
ACH68908.1_Uncultured ammonia-oxidizing archaeon clone SAT E3 (salt marsh sediment)
Candidatus Nitrosoarchaeum limnia SFB1
7 seq x 170 m
170m 1
AAZ81320.1_Uncultured crenarchaeote clone MB C130m 2 (Monterrey Bay, 30 m)
ABS81000.1_Uncultured crenarchaeote clone S33-A-11 (Changjiang Estuary, sediment)
170m D1
17 seq x 170 m
CBL51858.1_Uncultured archaeon clone NH6 (Lake Llebreta, Spain)
CBL51827.1_Uncultured archaeon clone NA2 (Lake Llebreta, Spain)
CBL51813.1_Uncultured archaeon clone NC12 (Lake Llebreta, Spain)
ACD67596.1_Uncultured crenarchaeote clone LCR AOA (rhizosphere, Lake Grane)
AEN04471.1_Candidatus Nitrosotalea devanaterra
ABQ01435.1_Uncultured crenarchaeote clone 9A2 (forest soil)
ABY83788.1_Candidatus Nitrosocaldus yellowstonii
CBY93673.1_Candidatus Nitrososphaera sp. EN76 vienensis (garden soil)
ABY77595.1_Candidatus Nitrososphaera gargensis (hot spring)
ABI21673.1_Uncultured crenarchaeote clone E09 (vapor cave)
Water column B
'Deep water' marine clade
Water column A/Sediments
'Shallow water' marine clade
Freshwater/soils
soils

Figure 1

Figure 2

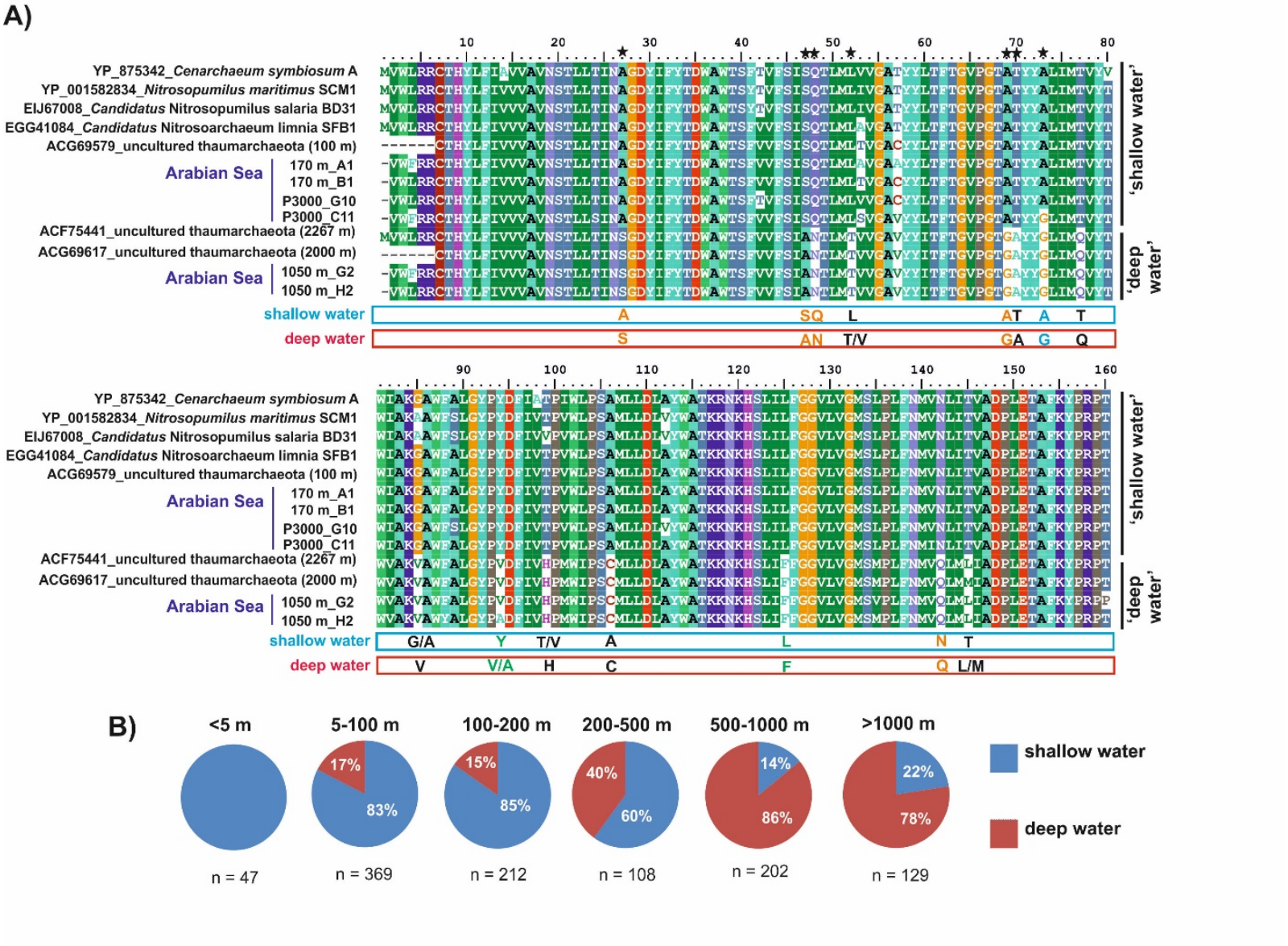

A)
YP_875342_Cenarchaeum symbiosum A
YP_001582834_Nitrosopumilus maritimus SCM1
EIJ67008_Candidatus Nitrosopumilus salaria BD31
EGG41084_Candidatus Nitrosoarchaeum limnia SFB1
ACG69579_uncultured thaumarchaeota (100 m)
Arabian Sea
170 m_A1
170 m_B1
P3000_G10
P3000_C11
ACF75441_uncultured thaumarchaeota (2267 m)
ACG69617_uncultured thaumarchaeota (2000 m)
Arabian Sea
1050 m_G2
1050 m_H2
'shallow water'
'deep water'
shallow water
deep water
B)
<5 m
5-100 m
100-200 m
200-500 m
500-1000 m
>1000 m
17%
83%
15%
85%
40%
60%
14%
86%
22%
78%
shallow water
deep water
n = 47
n = 369
n = 212
n = 108
n = 202
n = 129

Figure 3

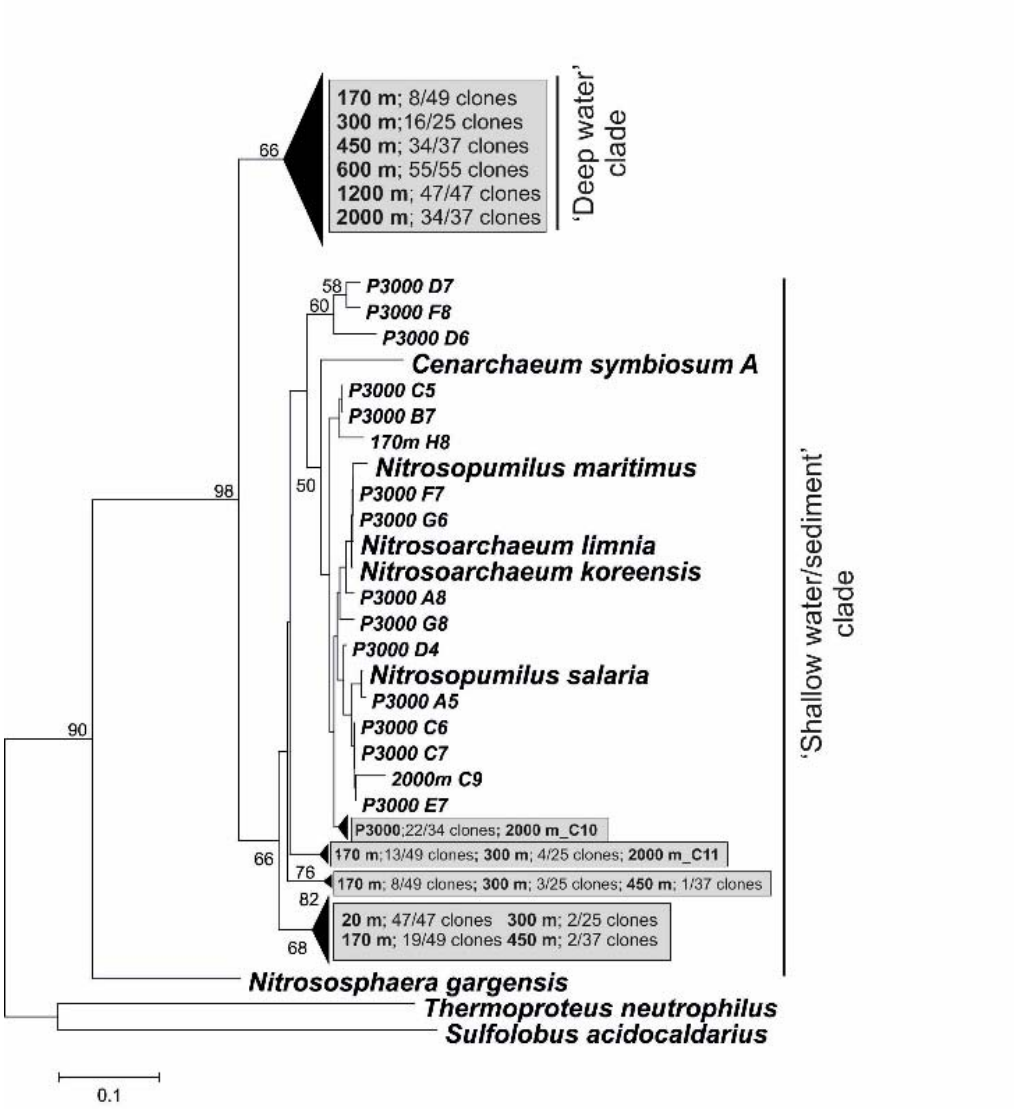

170 m; 8/49 clones
300 m;16/25 clones
450 m; 34/37 clones
600 m; 55/55 clones
1200 m; 47/47 clones
2000 m; 34/37 clones
'Deep water' clade
P3000 D7
P3000 F8
P3000 D6
Cenarchaeum symbiosum A
P3000 C5
P3000 B7
170m H8
Nitrosopumilus maritimus
P3000 F7
P3000 G6
Nitrosoarchaeum limnia
Nitrosoarchaeum koreensis
P3000 A8
P3000 G8
P3000 D4
Nitrosopumilus salaria
P3000 A5
P3000 C6
P3000 C7
2000m C9
P3000 E7
P3000;22/34 clones; 2000 m_C10
170 m;13/49 clones; 300 m; 4/25 clones; 2000 m_C11
170 m; 8/49 clones; 300 m; 3/25 clones; 450 m; 1/37 clones
20 m; 47/47 clones 300 m; 2/25 clones
170 m; 19/49 clones 450 m; 2/37 clones
'Shallow water/sediment' clade
Nitrososphaera gargensis
Thermoproteus neutrophilus
Sulfolobus acidocaldarius
66
58
60
98
50
90
66
76
82
68
0.1

Figure 4

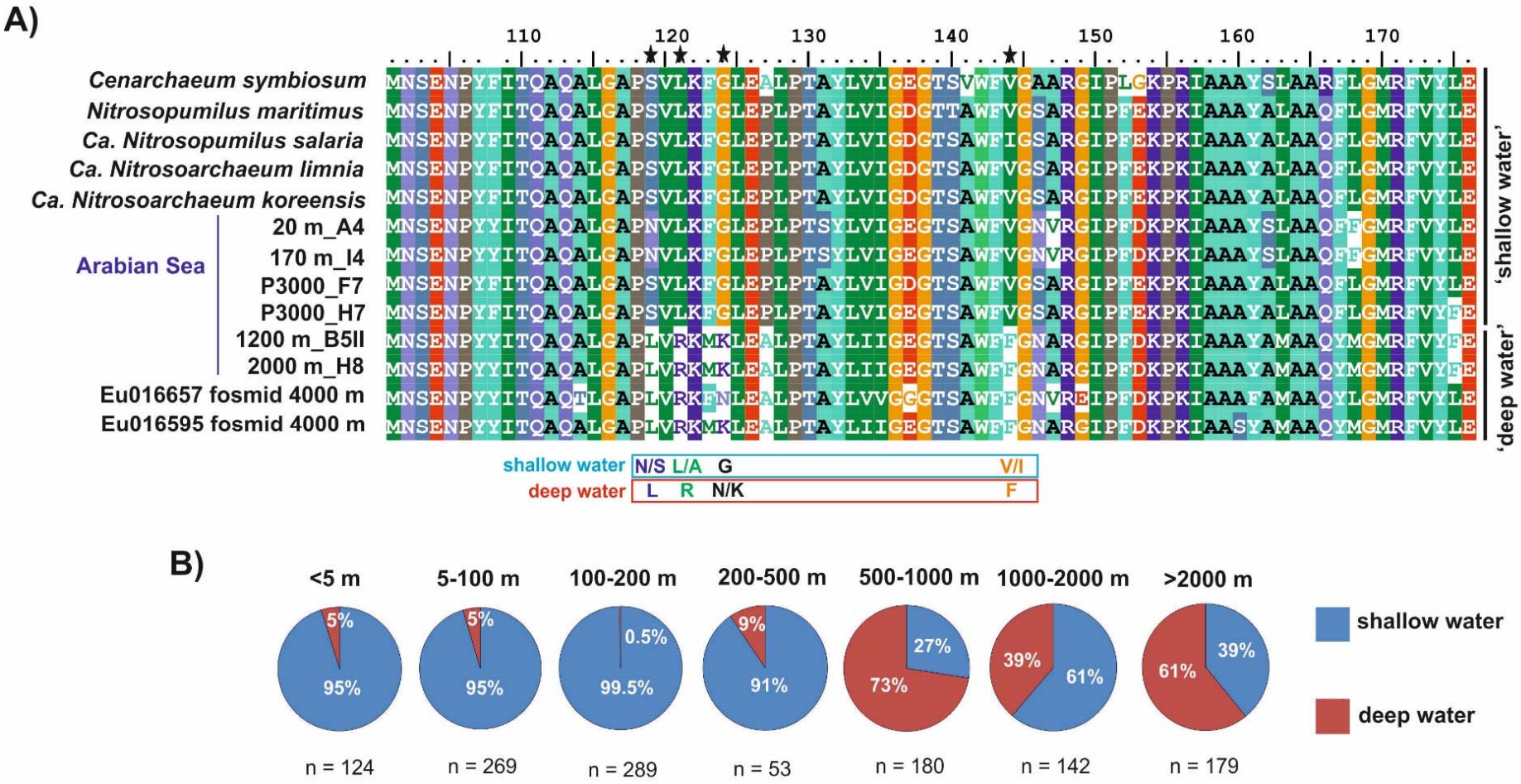

A)
110
120
130
140
150
160
170
Cenarchaeum symbiosum
Nitrosopumilus maritimus
Ca. Nitrosopumilus salaria
Ca. Nitrosoarchaeum limnia
Ca. Nitrosoarchaeum koreensis
Arabian Sea
20 m_A4
170 m_I4
P3000_F7
P3000_H7
1200 m_B5II
2000 m_H8
Eu016657 fosmid 4000 m
Eu016595 fosmid 4000 m
'shallow water'
'deep water'
shallow water N/S L/A G V/I
deep water L R N/K F
B)
<5 m
5-100 m
100-200 m
200-500 m
500-1000 m
1000-2000 m
>2000 m
5%
95%
5%
95%
0.5%
99.5%
9%
91%
27%
73%
39%
61%
61%
39%
shallow water
deep water
n = 124
n = 269
n = 289
n = 53
n = 180
n = 142
n = 179

Figure 5

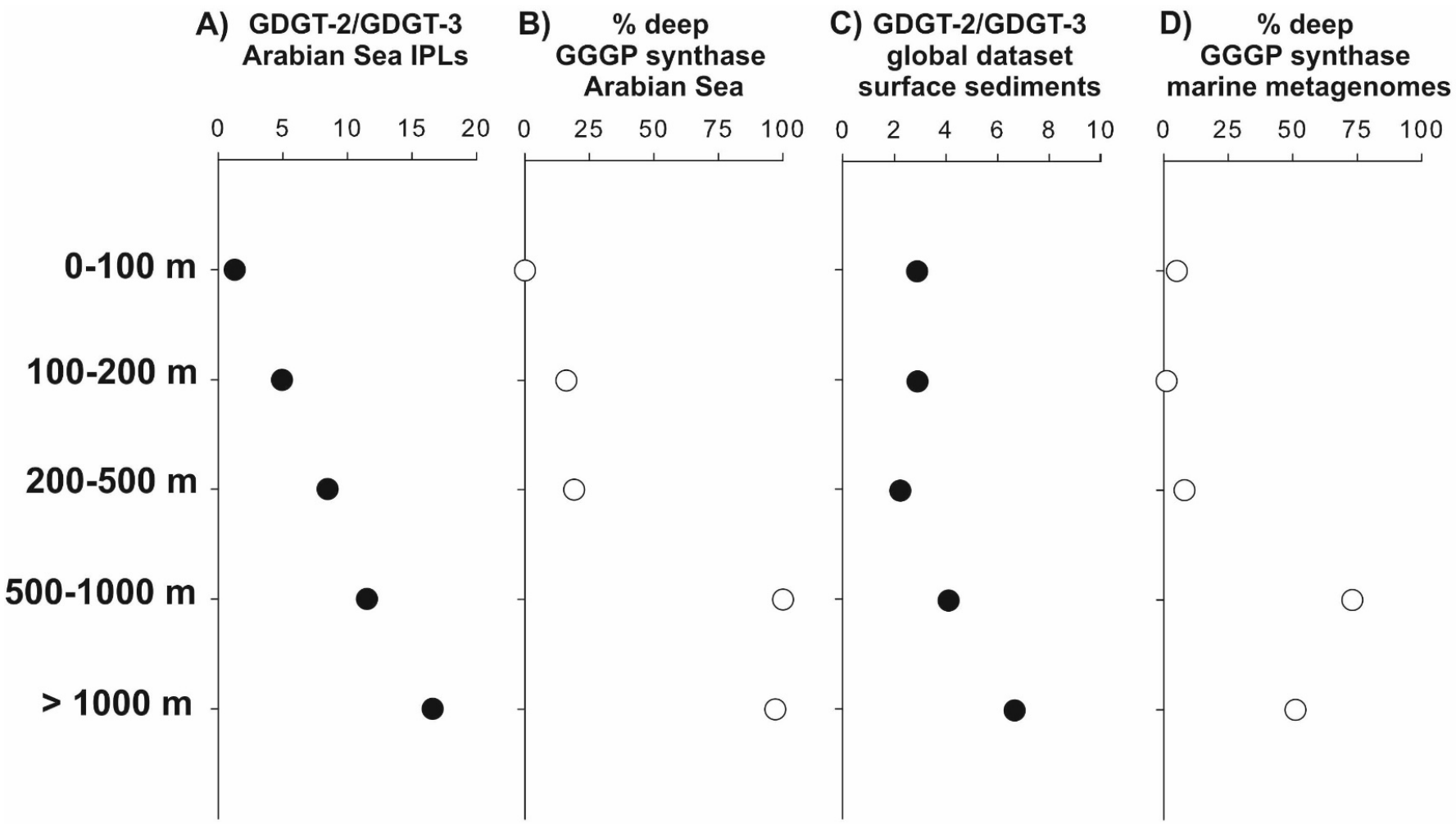

A) GDGT-2/GDGT-3
Arabian Sea IPLs
B) % deep
GGGP synthase
Arabian Sea
C) GDGT-2/GDGT-3
global dataset
surface sediments
D) % deep
GGGP synthase
marine metagenomes
0 5 10 15 20
0 25 50 75 100
0 2 4 6 8 10
0 25 50 75 100
0-100 m
100-200 m
200-500 m
500-1000 m
> 1000 m

Figure S1

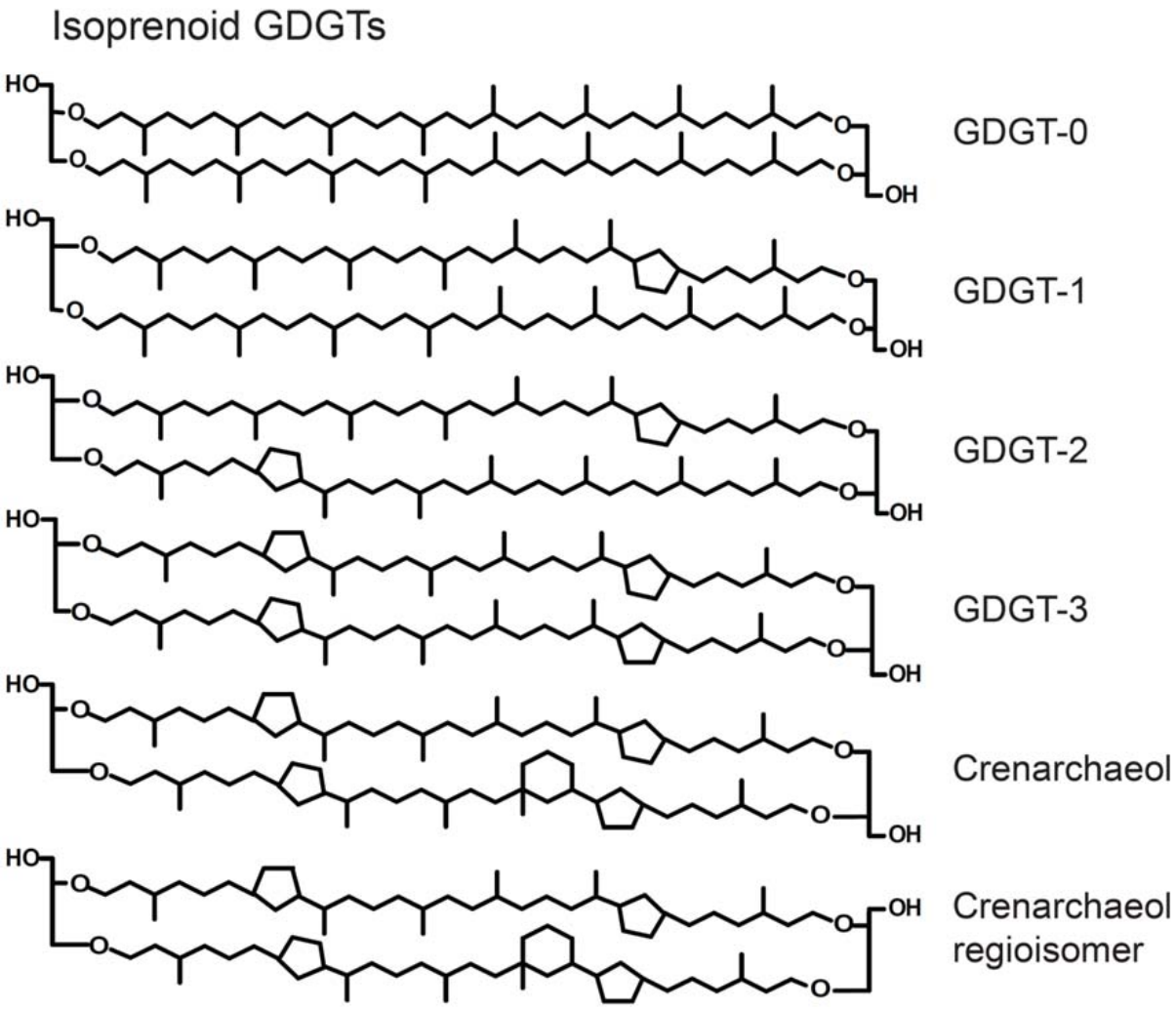

Isoprenoid GDGTs
HO
O
OH
GDGT-0
GDGT-1
GDGT-2
GDGT-3
Crenarchaeol
Crenarchaeol
regioisomer

Figure S2

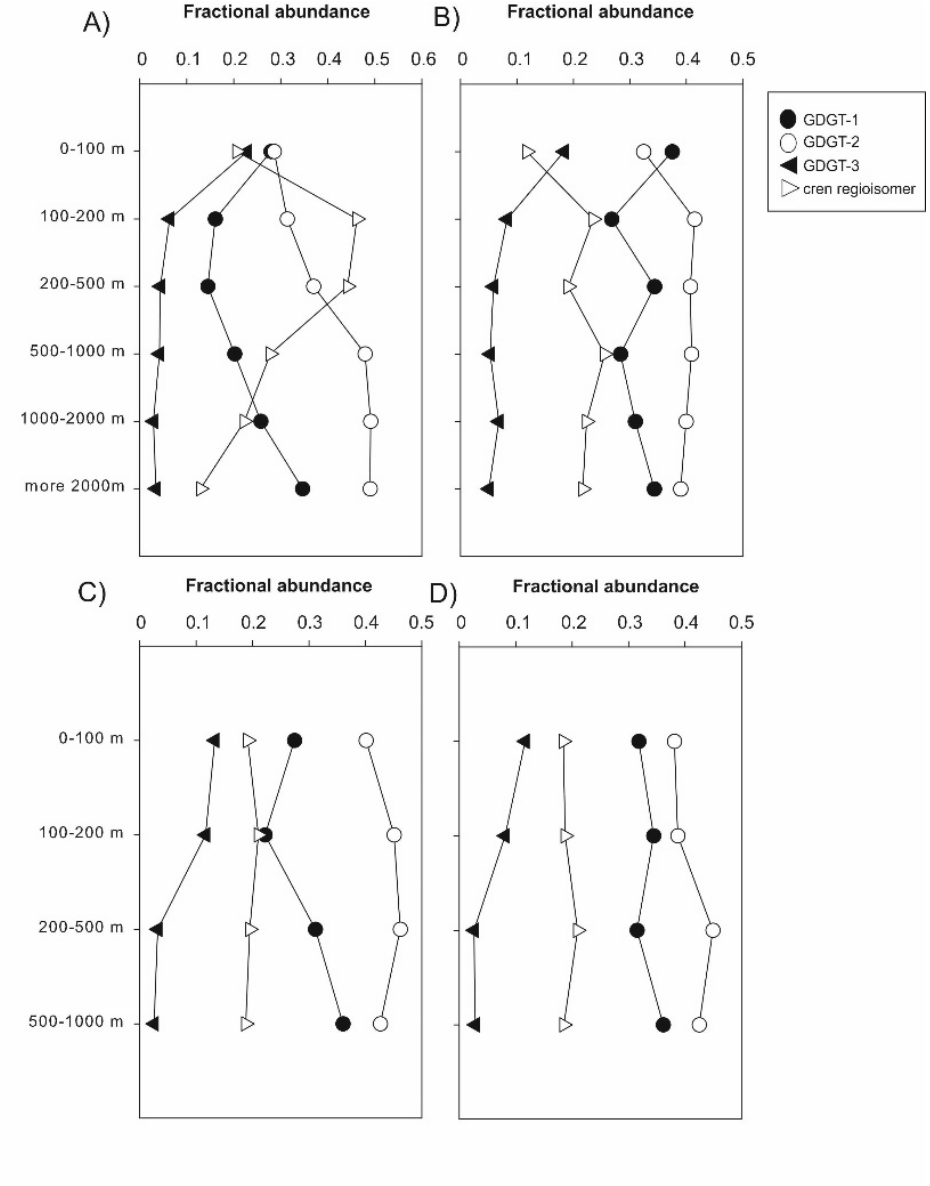

A)
Fractional abundance
0 0.1 0.2 0.3 0.4 0.5 0.6
B)
Fractional abundance
0 0.1 0.2 0.3 0.4 0.5
GDGT-1
GDGT-2
GDGT-3
cren regioisomer
0-100 m
100-200 m
200-500 m
500-1000 m
1000-2000 m
more 2000m
C)
Fractional abundance
0 0.1 0.2 0.3 0.4 0.5
D)
Fractional abundance
0 0.1 0.2 0.3 0.4 0.5
0-100 m
100-200 m
200-500 m
500-1000 m

Figure S3

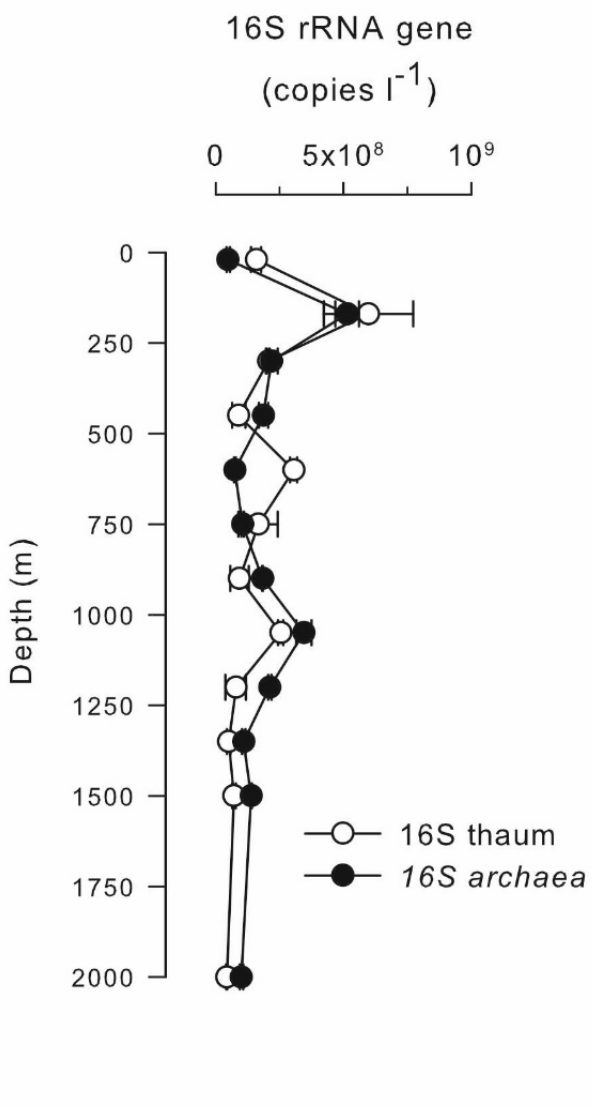
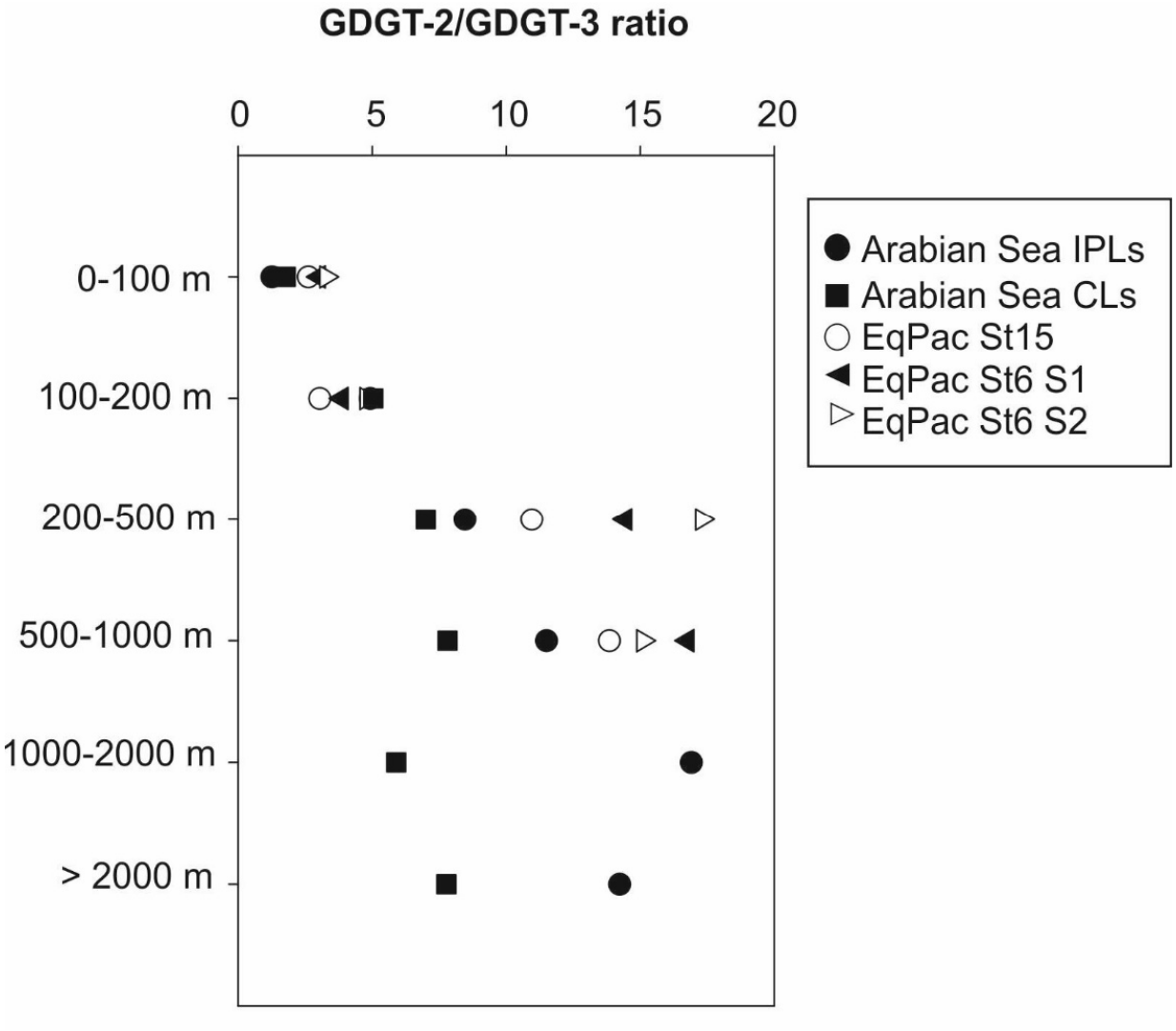

16S rRNA gene
(copies l-1)
0
5x10^8
10^9
Depth (m)
0
250
500
750
1000
1250
1500
1750
2000
16S thaum
16S archaea

Figure S4


GDGT-2/GDGT-3 ratio
0
5
10
15
20
0-100 m
100-200 m
200-500 m
500-1000 m
1000-2000 m
> 2000 m
Arabian Sea IPLs
Arabian Sea CLs
EqPac St15
EqPac St6 S1
EqPac St6 S2