environmental microbiology

Environmental Microbiology (2011) doi:10.1111/j.1462-2920.2011.02628.x

Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine

Benjamin J. Tully,[1]* William C. Nelson[1,2] and John F. Heidelberg[1,2]

[1]*Department of Biological Sciences, David and Dana Dornsife College of Letters, Arts and Sciences, University of Southern California, 3616 Trousdale Parkway, Los Angeles, CA 90089, USA.*

[2]*Wrigley Institute for Environmental Studies, 3616 Trousdale Parkway, AHF 410, University of Southern California, Los Angeles, CA 90089, USA.*

Summary

***Thaumarchaea*, which represent as much as 20% of prokaryotic biomass in the open ocean, have been linked to environmentally relevant biogeochemical processes, such as ammonia oxidation (nitrification) and inorganic carbon fixation. We have used culture-independent methods to study this group because current cultivation limitations have proved a hindrance in studying these organisms. From a metagenomic data set obtained from surface waters from the Gulf of Maine, we have identified 36 111 sequence reads (containing 30 Mbp) likely derived from environmental planktonic *Thaumarchaea*. Metabolic analysis of the raw sequences and assemblies identified copies of the catalytic subunit required in aerobic ammonia oxidation. In addition, genes that comprise a nearly complete carbon assimilation pathway in the form of the 3-hyroxypropionate/4-hydroxybutyrate cycle were identified. Comparative genomics contrasting the putative environmental thaumarchaeal sequences and '*Candidatus* Nitrosopumilus maritimus SCM1' revealed a number of genomic islands absent in the Gulf of Maine population. Analysis of these genomic islands revealed an integrase-associated island also found in distantly related microbial species, variations in the abundance of genes predicted to be important in thaumarchaeal respiratory chain, and the absence of a high-affinity phosphate uptake operon. Analysis of the underlying sequence diversity suggests the presence of at least two dominant environmental populations. Attempts to assemble complete environmental genomes were unsuccessful, but analysis of scaffolds revealed two diverging populations, including a thaumarchaeal-related scaffold with the full urease operon. Ultimately, the analysis revealed a number of insights into the metabolic potential of a predominantly uncultivated lineage of organisms. The predicted functions in the thaumarchaeal metagenomic sequences are directly supported by historic measurements of nutrient concentrations and provide new avenues of research in regards to understanding the role *Thaumarchaea* play in the environment.**

Received 31 March, 2011; accepted 25 September, 2011. *For correspondence. E-mail tully.bj@gmail.com; Tel. (+1) 213 740 4748; Fax (+1) 213 740 8132.

Introduction

Thaumarchaea comprise a large percentage of the marine planktonic microbe community, with as much as 40–50% of all prokaryotes in the deep ocean, and about 20% of all marine planktonic prokaryotes (Karner *et al*., 2001). All known members of the marine 1.1a group *Thaumarchaea* are obligately chemoautotrophic, deriving energy through aerobic ammonia oxidation (Hallam *et al*., 2006a; de la Torre *et al*., 2008; Hatzenpichler *et al*., 2008; Walker *et al*., 2010; Blainey *et al*., 2011) and fixing bicarbonate through the 3-hydroxypropionate/4-hydroxybutyrate cycle (Berg *et al*., 2007; 2010; Walker *et al*., 2010). Recent work has shown that an isolate of the 1.1b group *Thaumarchaea*, *Nitrososphaera viennensis*, has enhanced growth when grown in the presence of pyruvate, but still generates a majority of cellular carbon through autotrophy (Tourna *et al*., 2011). To date, five species from this phylum have had their genomes fully sequenced – a symbiont of the sponge *Axinella mexicana*, '*Candidatus* Cenarchaeum symbiosum A' (Hallam *et al*., 2006b); a marine isolate from the Seattle aquarium, '*Candidatus* Nitrosopumilus maritimus SCM1' (Walker *et al*., 2010); a moderately thermophilic enrichment culture, '*Candidatus* Nitrosphaera gargensis' (Hatzenpichler *et al*., 2008); a thermophilic isolate from Yellowstone National Park, '*Candidatus* Nitrosocaldus yellowstonii' (de la Torre *et al*., 2008); and a low-salinity enrichment culture from San Francisco Bay, '*Candidatus* Nitrosoarchaeum limnia SFB1' (Blainey *et al*., 2011). Additionally, there are a number of environmental sequences (Béjá *et al*., 2002; Lopez-Garcia *et al*., 2004; Konstantinidis and DeLong, 2008), and an increasing

number of isolates and enrichment cultures of *Thaumarchaea* undergoing sequencing and annotation. Extensive molecular ecology has demonstrated the global importance of the *Thaumarchaea*; their 16S rRNA gene sequence and the gene of ammonia monooxygenase subunit A (*amoA*) have been amplified from various habitats around the world (Venter *et al*., 2004; Francis *et al*., 2005; Park *et al*., 2006; Mincer *et al*., 2007; Bernhard *et al*., 2010; Church *et al*., 2010; Labrenz *et al*., 2010; Molina *et al*., 2010; Santoro *et al*., 2010). In the oceans, *Thaumarchaea* play a critical role in the carbon and nitrogen cycle; calculations of *in situ* ammonia oxidation and carbon fixation estimate that this group of organisms is capable of producing enough nitrite/nitrate to account for all 'new' nitrogen in the upper ocean and 1% of total global carbon fixation respectively (Ingalls *et al*., 2006; Berg *et al*., 2007). Despite their importance to the global marine ecosystem, the isolation and sequencing of a planktonic marine thaumarcheaote has yet to be achieved.

Without a planktonic marine isolate, many of the assumptions regarding the genomic potential and physiology of marine planktonic thaumarchaea are derived from molecular studies of 16S rRNA gene sequences and *amoA* genes. Previous studies have shown that the *Thaumarchaea* in the planktonic marine environment are capable of fixing carbon (Wuchter *et al*., 2003; Ingalls *et al*., 2006), yet molecular studies have not been used to explore the presence or abundance of genes implicated in thaumarchaeal carbon fixation (Berg *et al*., 2007; Walker *et al*., 2010). Furthermore, it can only be assumed that the current marine genomes and isolates are good approximations of genomic diversity and physiology in the planktonic marine environment, since '*Ca*. C. symbiosum A' is an obligate symbiont and '*Ca*. N. maritimus SCM1' was isolated from an artificial environment.

As an alternative to isolation, we have analysed thaumarchaeal sequences identified within a large-scale marine metagenome derived from summer and winter samplings from the Gulf of Maine (GOM), an environment with high abundance of marine thaumarchaea. From the data set of 2.4 Gbp of sequence, we were able to identify (bin) 36 111 reads (30 Mbp) with putative thaumarchaeal origin. Key genes required for ammonia oxidation and carbon fixation were identified in the putative thaumarchaeal bin and environmentally derived sequences were compared with '*Ca*. N. maritimus SCM1' (Walker *et al*., 2010) to explore the population and functional diversity of the GOM *Thaumarchaea*.

Results and discussion

Environmental sequencing and binning

An initial assembly of the GOM metagenomic data set was screened for contigs and scaffolds (clone-linked contigs) that originated from planktonic thaumarchaea. The initial GOM assembly produced 117 673 scaffolds comprised of 124 610 contigs, of which 738 scaffolds (0.59% of the total scaffolds) and 3662 contigs (2.9% of the total contigs) appeared thaumarchaeal-like. As the scaffolds are larger and thus contain more phylogenetic information, initial work to support the binning process was performed on the scaffolds. A comparison of the tetranucleotide frequency of these scaffolds to the '*Ca*. N. maritimus SCM1' genome yielded a z-score correlation of 0.93, supporting our assignment of the scaffolds to the *Thaumarchaea* (Teeling *et al*., 2004). Furthermore, two scaffolds were identified that contain the small subunit rRNA genes. The two corresponding 16S rRNA genes are nearly identical [99.5% nucleic acid identity (NAID)]. The divergence between the 16S rRNA sequences is much greater than the 0.001% error rate estimated for Sanger-derived sequences processed using the phred base-calling program (Ewing and Green, 1998), supporting that the scaffolds represent two distinct groups. Additionally, they have high identity to the '*Ca*. N. maritimus SCM1' 16S rRNA gene (98.7% and 98.6% NAID). While it is unknown to what degree *Thaumarchaea* follow the microbial species definition [> 97% 16S rRNA gene sequence identity (Hagström *et al*., 2000)], this similarity suggests that the environmental data set contains at least two closely related organisms which may represent species of the genus *Nitrosopumilus*.

In total, the thaumarchaeal metagenomic assemblies (scaffolds and contigs) are composed of 36 111 sequences, averaging 837 bp in length, containing a total of 30.2 Mbp. Almost all of these sequences were derived from the winter libraries (99.3%). For the three libraries generated from the winter samples (see *Experimental procedures*), the average percentage of thaumarchaeal-like reads per library is 2.81% (range, 2.11–3.61%) (Table 1), suggesting that the *Thaumarchaea* may make-up a similar percentage of the planktonic prokaryotic community at all three sites. About two-thirds (66.3%) of the assemblies are a composite of sequences from at least two different GOM sites indicating that there is an overlap in the populations between the sample sites.

Ammonia oxidation and carbon fixation

The gene content of the GOM sequences suggests that, like other members of the *Thaumarchaea*, the populations in the GOM are capable of chemoautotrophy. Functional annotation shows that the ammonia monooxygenase subunits (*amoA*, *amoB* and *amoC*), ammonium permease (*amt*), the urease operon, urease transporter, a putative nitrite reductase (*nirK*) and a putative nitric-oxide reductase Q (*norQ*) are all present within the metagenomic sequences (Table 2). There was no evidence of hydroxy-

Table 1. Sample location and sequencing effort.

Site name	Date sampled	Latitude	Longitude	Temperature (°C)	Salinity (ppt)	Number of sequences	Per cent of sequences related to *Thaumarchaea*	Insert size range (kb)
GOM03	25-Jan-06	42°46′9″N	68°40′8″W	6.4	33	453 805	2.11	3–5/2–3
GOM04	27-Jan-06	44°07′5″N	67°58′3″W	5.1	32.2	957 737	2.70	4–6
GOM06	30-Jan-06	41°28′7″N	69°6′0″W	6.2	32.6	10 040	3.61	4–6
GOM12	25-Aug-06	41°08′6″N	66°53′3″W	17.4	32	470 591	0.01	6–8
GOM13	28-Aug-06	43°23′3″N	67°41′9″W	16.6	32	925 793	0.02	6–8/8–10
GOM14	29-Aug-06	42°21′6″N	69°23′8″W	19.6	31.3	9 728	0.00	6–8

lamine oxidoreductase, which catalyses the second step of aerobic ammonia oxidation in bacteria; however, three multicopper oxidases (MCO) were identified which could convert hydroxylamine/nitroxyl to nitrite (Walker *et al*., 2010). Furthermore, like '*Ca*. N. maritimus SCM1', a number of transporters were detected which are capable of amino acid uptake. The presence of this suite of genes indicates that the GOM planktonic thaumarchaea are putatively involved in generating energy through the oxidation of ammonia and harvesting amino acids from the environment. The presence of both heterotrophic and chemoautotrophic genes in the environment suggests a potential for mixotrophy, but has not be seen in experiments with the thaumarchaeal isolates (de la Torre *et al*., 2008; Walker *et al*., 2010; Blainey *et al*., 2011).

Our observations suggest the possibility that only a minority of the planktonic thaumarchaeal population in the GOM is capable of using urea as a nitrogen source. For the GOM metagenome, the genes, which comprise the urease operon, recruit fewer homologous sequences (1–10 sequences) than the genes that encode ammonium transporters (135 sequences) and the ammonia oxidation pathway (45–79 sequences) (Table 2). The number of metagenomic sequences homologous to any particular gene potentially provides information regarding the underlying abundance of that gene within an environmental population, although this correlation breaks down if there are varying copy numbers between genomes or if cloning bias causes the genes to be undersampled. In terms of environmental support for a thaumarchaeal subpopulation capable of using urea as a nitrogen source, during the stratified summer months, urea concentration exceeds the ammonium concentration in the surface water, generating a large pool of urea that could be used by organisms (Christensen *et al*., 1996; Dyhrman and Anderson, 2003).

All isolated members of the *Thaumarchaea* have the ability to grow autotrophically using the

Table 2. Sequence recruitment for genes related to the ammonia oxidation pathway.

Gene name	Gene source genome	Range of %ID of recruited sequences	Total number of sequence
Urease α subunit	*Cenarchaeum symbiosum* A	52–79%	6
Urease β subunit	*Cenarchaeum symbiosum* A	64%	1
Urease γ subunit	*Cenarchaeum symbiosum* A	69–73%	2
Urease Accessory E	*Cenarchaeum symbiosum* A	44–57%	3
Urease Accessory F	*Cenarchaeum symbiosum* A	35–46%	5
Urease Accessory G	*Cenarchaeum symbiosum* A	56–74%	6
Urease Accessory H	*Cenarchaeum symbiosum* A	41–58%	10
Urea transporter	*Cenarchaeum symbiosum* A	45–83%	9
Hydroxylamine oxidoreductase	*Desulfobacterium autotrophicum* HRM2	–	0
Nitrate reductase [*nar*]	*Haloarcula marismortui* ATCC 43049	–	0
Ferrodoxin nitrite/sulfite reductase [*nirA*]	*Cenarchaeum symbiosum* A	21–32%	79
Ferrodoxin nitrite reductase NAD(P)H [*nirB*]	*Vibrio fischeri* MJ11	26–27%	4
Formate-dependent nitrite reductase [*nrfA*]	*Vibrio fischeri* MJ11	–	0
Nitrite reductase [*nirK*]	*Haloarcula marismortui* ATCC 43049	26–40%	25
Nitric-oxide reductase			
norB	*Haloarcula marismortui* ATCC 43049	–	0
norC	*Silicibacter pomeroyi* DSS-3	–	0
norQ	*Cenarchaeum symbiosum* A	27–33%	70
norD	*Silicibacter pomeroyi* DSS-3	–	0
Nitrous-oxide reductase [*norZ*]	*Geobacillus thermodenitrificans* NG80-2	–	0
Ammonium transporter	*Cenarchaeum symbiosum* A	33–82%	135
ammonia monooxygenase subunit A	*Cenarchaeum symbiosum* A	72–96%	63
ammonia monooxygenase subunit B	*Cenarchaeum symbiosum* A	58–92%	46
ammonia monooxygenase subunit C	*Cenarchaeum symbiosum* A	69–98%	45

Fig. 1. 3-Hydroxypropionate/4-Hydroxybutyrate cycle. Green boxes indicate genes with reads that recruit at ≥ 30% amino acid identity. Yellow boxes indicate genes with reads that recruit at ≥ 20%/< 30% amino acid identity. Red boxes indicate genes without recruited reads. Numbers with each box correspond to Enzyme Commission (EC) numbers. Enzymes represented by each box: (1) acetyl-CoA carboxylase; (2) malonyl-CoA reductase (NADPH); (3) malonate semialdehyde reductase (NADPH); (4) 3-hydroxypropionyl-CoA synthetase (AMP-forming); (5) 3-hydroxypropionyl-CoA dehydratase; (6) acryloyl-CoA reductase (NADPH); (7) propionyl-CoA carboxylase; (8) methylmalonyl-CoA epimerase; (9) methylmalonyl-CoA mutase; (10) succinyl-CoA reductase (NADPH); (11) succinate semialdehyde reductase (NADPH); (12) 4-hydroxybutyryl-CoA synthetase (AMP-forming); (13) 4-hydroxybutyryl-CoA dehydratase; (14) crotonyl-CoA hydratase; (15) 3-hydroxybutyryl-CoA dehydrogenase (NAD+); (16) acetoacetyl-CoA b-ketothiolase.

3-hydroxypropionate/4-hydroxybutyrate cycle (Berg *et al*., 2007; 2010; Walker *et al*., 2010). Using the sequences for each step in the pathway as identified in *Metallosphaera sedula* DSM 5348, all of the genes were present in the putative GOM *Thaumarchaea*, except for malonate semialdehyde reductase and ^{4}OH-butyryl-CoA synthetase (Fig. 1; Table S1) (Berg *et al*., 2007). Fifty-six sequences from the planktonic thaumarchaea bin appear to be homologous to the malonate semialdehyde reductase of marine γ-*Proteobacterium* HTCC2080, with the highest scoring sequence in this search (Read ID: 1106160064005) having only 27% amino acid identity (AAID). However, no putative CDS from the GOM metagenome, '*Ca*. C. symbiosum A' and '*Ca.* N. maritimus SCM1' have significant similarity (*E*-value ≤ 10^{-3}) to the ^{4}OH-butyryl-CoA synthetase either from *M. sedula* DSM 5348 (Msed_1456) or from *Sulfolobus tokodaii* 7 (ST0783) (Alber *et al*., 2008). It is possible that the malonate semialdehyde reductase and the ^{4}OH-butyryl-CoA synthetase functions have been replaced by marine thaumarchaeal-specific enzymes. Several putative CDS were identified with similarity to a locus identified in '*Ca.* N. maritimus SCM1' as being ^{4}OH-butyryl-CoA synthetase (Nmar_206) (Walker *et al*., 2010); however, we found no evidence, either experimental or computational, to support or refute this annotation. The thaumarchaeal bin was also searched for key genes in the other carbon fixation pathways, and no orthologues were detected. Thus, despite the incomplete gene set identified, it is likely that these planktonic thaumarchaea are able to fix bicarbonate through the ^{3}OH-propionate/^{4}OH-butyrate cycle, though, as has been shown in the 1.1b group *Thaumarchaea*, some marine thaumarchaea may experience enhanced growth using some form of mixotrophy (Tourna *et al*., 2011).

Comparative genomics to 'Ca. N. maritimus SCM1'

The 36 111 sequences that composed the thaumarchaeal bin were recruited against the '*Ca.* N. maritimus SCM1' genome. A total of 33 037 sequences (91.49%) aligned. In comparison, all the reads generated from the Global Ocean Survey (GOS) phases I and II (14 274 550 reads), only 31 504 reads recruit to the '*Ca.* N. maritimus SCM1' genome. For the GOM thaumarchaeal bin, the average pairwise nucleotide identity to the '*Ca.* N. maritimus SCM1' genome was 76.6% and average coverage depth was 17.3× (range, 0–54; standard deviation, 12.0). The GOM metagenome more than doubles the depth of coverage available for analysis in comparing the '*Ca.* N. maritimus SCM1' genome to environmental organisms.

The J. Craig Venter Institute Annotation Pipeline identified 44 701 putative coding sequences (CDSs) in the thaumarchaeal bin. A vast majority of the putative CDS

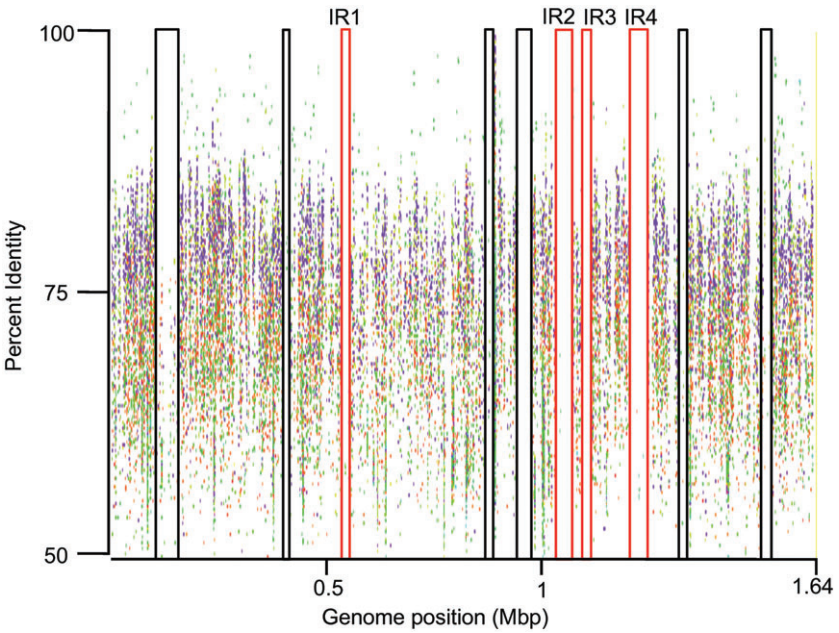

Fig. 2. Recruitment plot of GOM metagenome and GOS Phase 1 metagenomes against the '*Ca.* N. maritimus SCM1' reference genome. Red boxes indicate integrase regions (IR), genomic islands with an annotated integrase gene. Black boxes indicate large islands present in both GOM and GOS. Purple coloured reads are from the various GOM libraries. GOS metagenomes are represented by all other colours.

(87.6%) had highest similarity to genes from the '*Ca.* N. maritimus SCM1' genome. An additional, 10.0% of the putative CDS had best similarity to genes from other organisms, but still had similarity to a gene from '*Ca.* N. maritimus SCM1'. These CDSs represent genetic material that is more divergent from the *Ca. N. maritimus* SCM1 orthologues. Additionally, 1.4% of the putative CDSs only had similarity to genes from other marine thaumarchaeal sources (see *Experimental procedures*) and 1.0% only had similarity to genes from non-thaumarchaeal sequences. These CDSs may represent genes with a deep thaumarchaeal lineage that were lost from '*Ca.* N. maritimus SCM1', or may represent genetic material that was horizontally transferred into the GOM population.

There were 69 regions in the '*Ca.* N. maritimus SCM1' genome for which no metagenomic sequences from the GOM were recruited. Most of the regions (43) were $\leq$ 2 kb in length. These gaps could be due to the random distribution of coverage common in genomic shotgun sequencing or due to genomic variation between '*Ca.* N. maritimus SCM1' and the GOM *Thaumarchaea*. Of the 26 coverage gaps that were $>$ 2 kb in length, 16 also fail to recruit sequences from the GOS phase I metagenomes (Fig. 2; Table S2). Although we do not have sufficient sequence depth to ensure complete coverage of the environmental thaumarcheaote, the length of these regions and the fact that independent samples show similar patterns of absence suggests that these gaps represent islands in '*Ca.* N. maritimus SCM1' that are not common in the planktonic thaumarchaea. It should be noted, however, that even large gaps found in reference genomes compared with metagenomic sequences have been found to be artefacts (Bhaya *et al.*, 2007).

Four of these large islands include annotated integrase genes, suggesting that '*Ca.* N. maritimus SCM1' acquired these regions through lateral gene transfer. Integrase region 1 (IR1) (Fig. 3) is relatively short (~16 kb) and contains 13 genes; seven are predicted hypothetical proteins, five genes encode restriction/modification functions, and one is the annotated integrase. The genes adjacent to the integrase gene are syntenic to regions found in the genomes of the *Euryarchaeota Ferroplasma acidarmanus* Fer1 (Allen *et al.*, 2007) and the *γ-Proteobacterium Klebsiella pneumoniae* 342 (Fouts *et al.*, 2008) (Fig. 3). The wide phylogenetic distance between these organisms suggests that this is likely a mobile genetic element that can move between the prokaryotic domains and classes, although it is possible that this region is a remnant of the last common ancestor of prokaryotes.

IR2–4 range in size from ~22 to 45 kb (18–37 genes). Many of these genes are annotated as hypothetical proteins. However, some of the genes with a functional annotation in IR2, IR4 and in other coverage gaps (Table S2) are related to the proposed thaumarchaeal respiratory chain for '*Ca.* N. maritimus SCM1' (Walker *et al.*, 2010), specifically, genes annotated as DsbA oxidoreductase, hypothesized to play a role in alleviating copper or nitric oxide toxicity, blue copper proteins (BCP), hypothesized to act as plastocyanin-like electron shuttles, and multicopper oxidases (MCO), predicted to mediate the second step of ammonia oxidation ($H_xNO_x \rightarrow NO_2^-$) (Walker *et al.*, 2010). In the '*Ca.* N. maritimus SCM1' genome, these genes are numerous – 11 DsbA oxidoreductases, 18 BCPs and 4 MCOs – but in the GOM and GOS metagenomes only a subset of these genes recruit reads – 6 DsbA oxidoreductase, 12 BCPs and 3 MCOs. Distinct

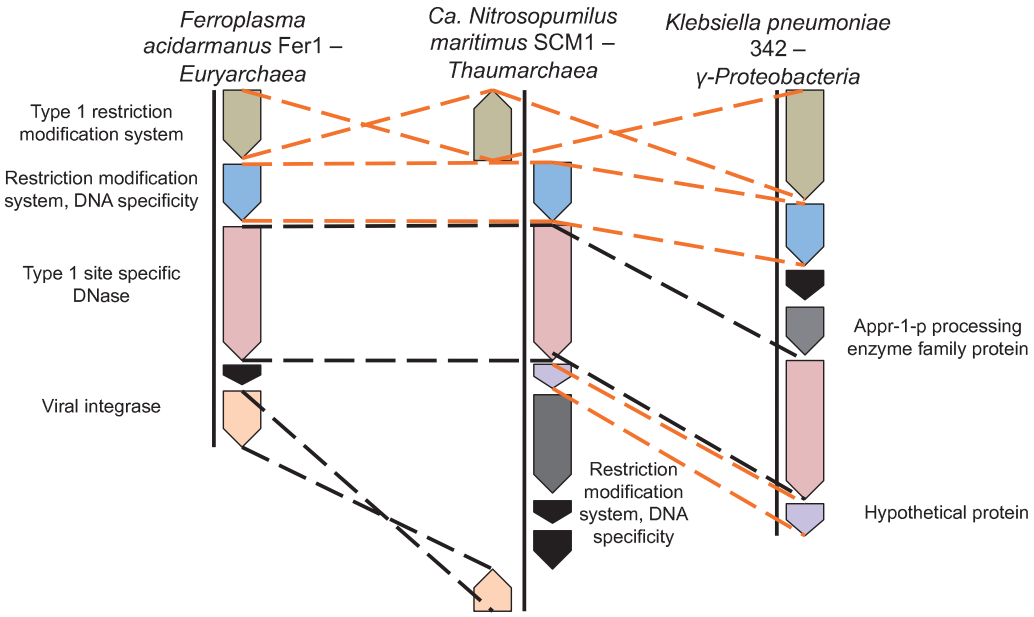

Fig. 3. Representation of integrase region 1 (IR1). Orange dashed lines indicate amino acid identity between genes > 30% identity. Black dashed lines indicate syntenic genes with the same PFAM assignment. Black arrows are hypothetical genes.

selective pressures due to differences in copper concentrations between the Seattle Aquarium and the marine environment might have resulted in the variation in gene content.

Another island includes a high-affinity phosphate uptake operon (*pstSCAB*). This region does not recruit sequences from the GOM metagenome (Island-22, Table S2), but does recruit sequences (at 58–75% NAID) from the GOS metagenomes from the Sargasso Sea, the east coast of the USA, the open ocean near Cuba, and the Smithsonian Tropical Research Institute in Panama (Rusch *et al.*, 2007). Loss of this operon in the GOM *Thaumarchaea* may be due to the higher phosphate levels found in the GOM. In the winter, phosphate levels range from ~0.5 to 1.0 μM is the surface waters of the GOM (World Ocean Database; http://www.nodc.noaa.gov), approximately 100 times more available phosphate than is found in the Sargasso Sea, where phosphate levels average 7.9 nM (Van Mooy *et al.*, 2006). These results highlight the importance of comparing isolate genomes and environmental metagenomes. The lack of the thaumarchaeal *pstSCAB* operon from the GOM metagenome would have likely gone unnoticed had the recruitment to '*Ca.* N. maritimus SCM1' not been performed. Such comparisons can thus increase our understanding of how the biology is responding to the system in both environments.

Metagenomic population diversity

The environmental shotgun sequencing allowed for an analysis of the underlying population. Due to the high abundance of microbial cells in the marine environment (~10^6 ml^{-1}), it is assumed that environmental sequence generated comes from a unique individual. The overall population diversity can then be examined at the individual level by comparing the sequence variations between reads. Variation in the thaumarchaeal population in the GOM was examined using genes from the ammonia oxidation (*amoA*) and carbon fixation pathways, as well as the archaeal DNA repair recombinases (*radA*). The GOM *amoA* sequence divergence ranged from 7.9% to 16.1% NAID to the '*Ca.* N. maritimus SCM1' sequence. These values are comparable to the divergence in the *amoA* database generated through environmental molecular studies (Venter *et al.*, 2004; Francis *et al.*, 2005; Park *et al.*, 2006; Mincer *et al.*, 2007; Bernhard *et al.*, 2010; Church *et al.*, 2010; Labrenz *et al.*, 2010; Molina *et al.*, 2010; Santoro *et al.*, 2010). The thaumarchaeal bin sequences were used to construct a maximum likelihood (ML) phylogenetic tree (Fig. 4). The sequences cluster into two groups, closely related to '*Ca.* N. maritimus SCM1'. This pattern of two dominant groups is present in the ML trees constructed for methylmalonyl-CoA mutase (Fig. 5), ^{4}OH-butyryl-CoA dehydratase (Fig. S1) (representing the carbon fixation pathway), and *radA* (Fig. S2). Maximum nucleotide divergence of these sequences is never greater than 20.2%, and all sequences are more closely related to '*Ca.* N. maritimus SCM1' than other *Archaea*. Sequences within each of the two groups are greater than 90% identical, suggesting that further assembly may result in two distinct dominant genomes.

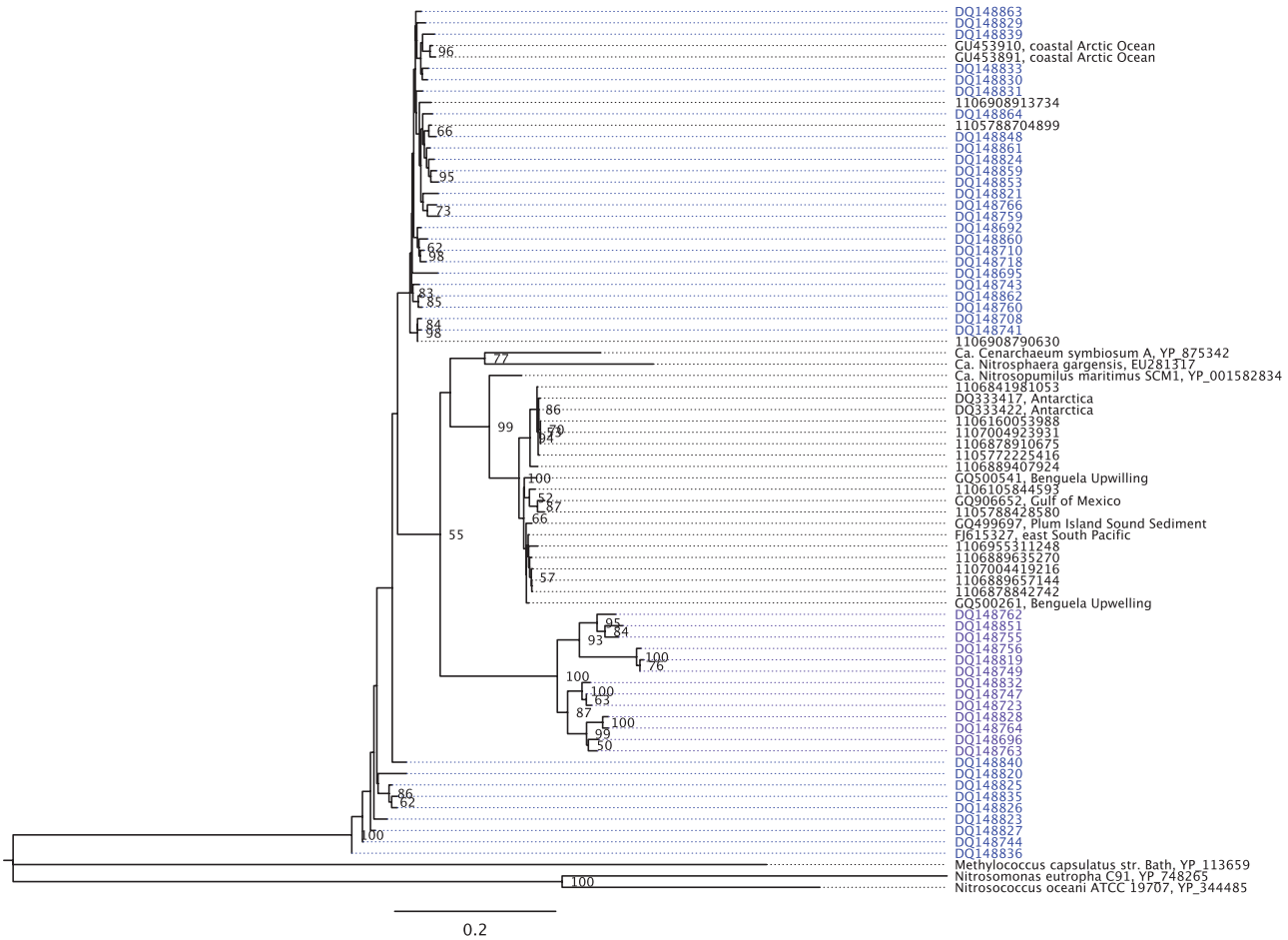

Fig. 4. Maximum likelihood tree (bootstrap: 1000) of a 422 bp region of the *amoA* gene from 16 GOM sequences (ID starting 110-), environmental sequences obtained from the GenBank nt database (black; Accession No.), sequences from 'water column group A' (Francis *et al*., 2005) (blue; Accession No.), and sequences from 'water column group B' (Francis *et al*., 2005) (purple; Accession No.).

Environmental *amoA* clones from various studies were added to the ML tree, including sequences from the GOS database and sequences that partitioned in to the 'water column clusters A and B' from Francis and colleagues (2005) (Fig. 4). Sequences from the GOM metagenome are interspersed with sequences from the east South Pacific (Molina *et al*., 2010), the Benguela Upwelling of the coast of Namibia (Moraru *et al*., 2010), Antarctica (Hallam *et al*., 2006a), particle associated communities in the Gulf of Mexico (B. Liu, Z. Huang, and C. Zhang, unpublished) and the coastal Arctic Ocean (Christman *et al*., 2011). The wide dispersal of *amoA* sequences closely related to those in the GOM metagenome may suggest that the specific thaumarchaeal gene is either under restrictive evolutionary pressure or rapidly dispersed through the environment. A similar pattern is seen in the methylmalonyl-CoA mutase ML tree, with sequences from Block Island (coastal north eastern USA), coastal waters off of South Carolina (south eastern USA) and Newcomb Bay (Antarctica) interspersed with GOM sequences (Fig. 5). Furthermore, all but two of *amoA* sequences retrieved from the GOM metagenome do not cluster with the depth-partitioned water column groups identified in Francis and colleagues (2005). The two new groups identified in the GOM metagenome may represent further diversity in the *amoA* gene.

Functional diversity

The underlying population diversity suggests that the *Thaumarchaea* present in the GOM are part of two dominant populations. The presence of two different, but closely related 16S rRNA sequences in the original GOM assembly further supports this hypothesis. Assuming each organism has a genome of a similar size to '*Ca.* N. maritimus SCM1' (~2 Mbp), and both variants have even representation in the community, there is sufficient sequence data in the thaumarchaeal bin to assemble each genome at ~7× coverage The 36 111 sequences within the bin were assembled using the Celera assem-

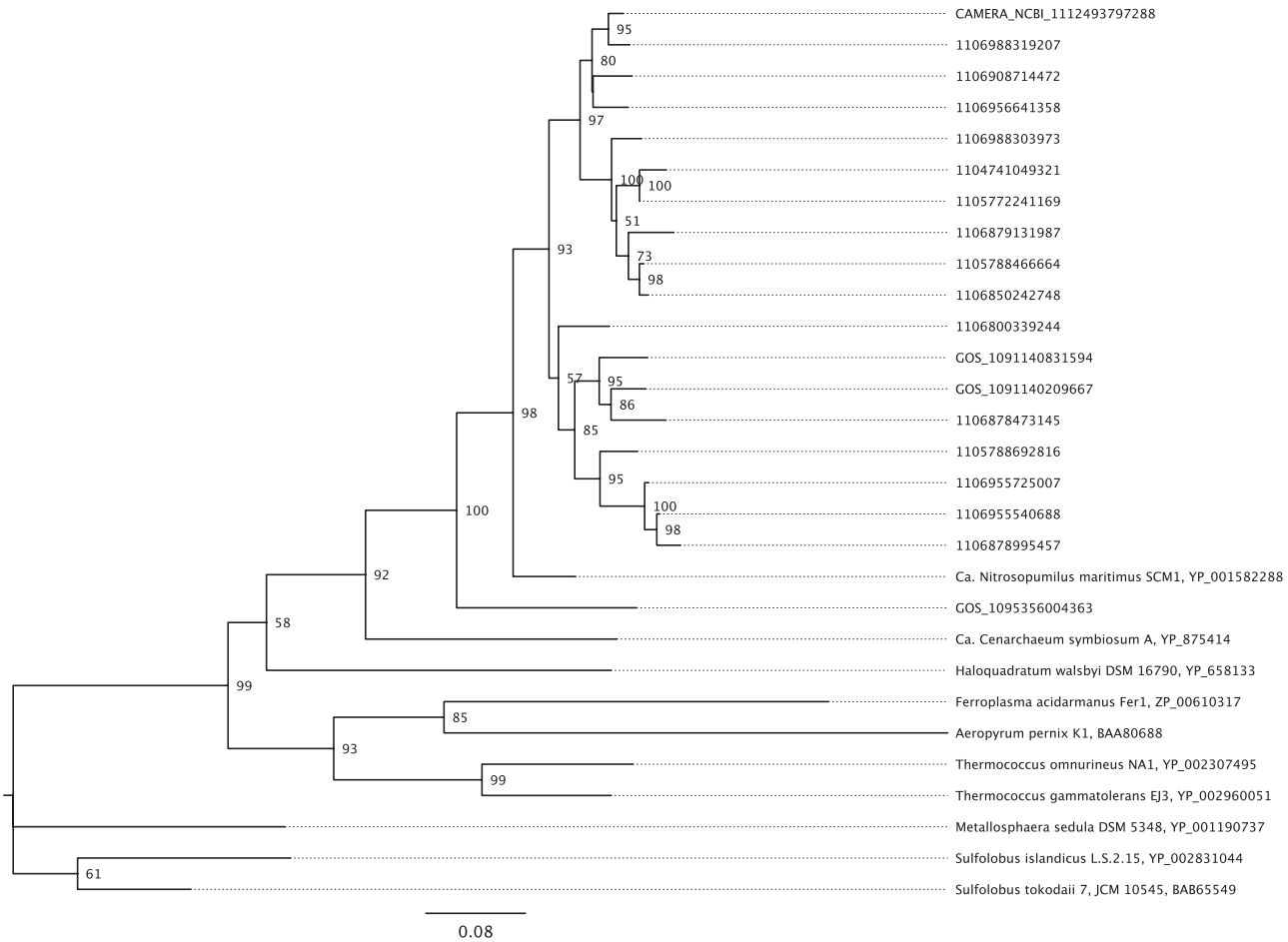

Fig. 5. Maximum likelihood tree (bootstrap: 1000) of a 388 bp region of the methylmalonyl-CoA synthetase gene from 15 GOM sequences (ID starting 110-) and environmental sequences obtained from the CAMERA database (CAMERA ID No.).

bler (Myers *et al*., 2000), with the parameter specification file modified to assemble microbial-sized genomes (see *Experimental procedures*). The assembly process generating 1584 scaffolds (longest = 162 170 bp) spanning over 6.4 Mbp of sequence. Only about 15% of the sequences did not assemble (5026 sequences) suggesting either that the underlying diversity of thaumarchaea in the GOM is more complex than indicated by our gene-centric analysis or that the binning process was inaccurate (e.g. lacking some thaumarchaeal sequences or containing non-thaumarchaeal sequences).

The scaffolds were used to compare genomic and functional variations between the two dominant GOM populations and '*Ca.* N. maritimus SCM1'. Specifically, the regions containing the ammonia monooxygenase subunits, ^{4}OH-butyryl-CoA dehydratase, and the environmental-based urease gene were analysed. For each of these regions (and for most of the other regions visually inspected), at least two 'long' (> 15 kb) scaffolds were present, putatively representing the two dominant populations in the environment, along with several smaller scaffolds (~3–7 kb). The 'long' scaffolds were further analysed because they offer the greatest opportunity to study variations in genomic synteny and single-nucleotide polymorphism (SNPs).

The alignments of the long scaffolds illustrated several different genomic rearrangements. Two 'long' environmental scaffolds contained the genes associated with ammonia oxidation. 'Scaffold A' (38 931 bp) had complete synteny with the region on the '*Ca.* N. maritimus SCM1' genome associated with ammonia oxidation (1 367 240–1 378 730 bp) except for two internal gaps (sequencing gaps determined during the assembly process) (Fig. 6). 'Scaffold B' (19 904 bp) possessed two insertions when compared with both Scaffold A and '*Ca.* N. maritimus SCM1'. The three putative CDS identified in the insertion sequences had ≥ 72% AAID to genes identified in '*Ca.* N. maritimus SCM1', '*Ca.* C. symbiosum A', and a fosmid clone derived from an uncultivated deep-sea *Thaumarchaea* (HF4000_APKG3E18) (Konstantinidis and DeLong, 2008). *amoB* and *amoC* on 'Scaffold A' had 89.6% and 88.5% NAID to '*Ca.* N. maritimus SCM1'.

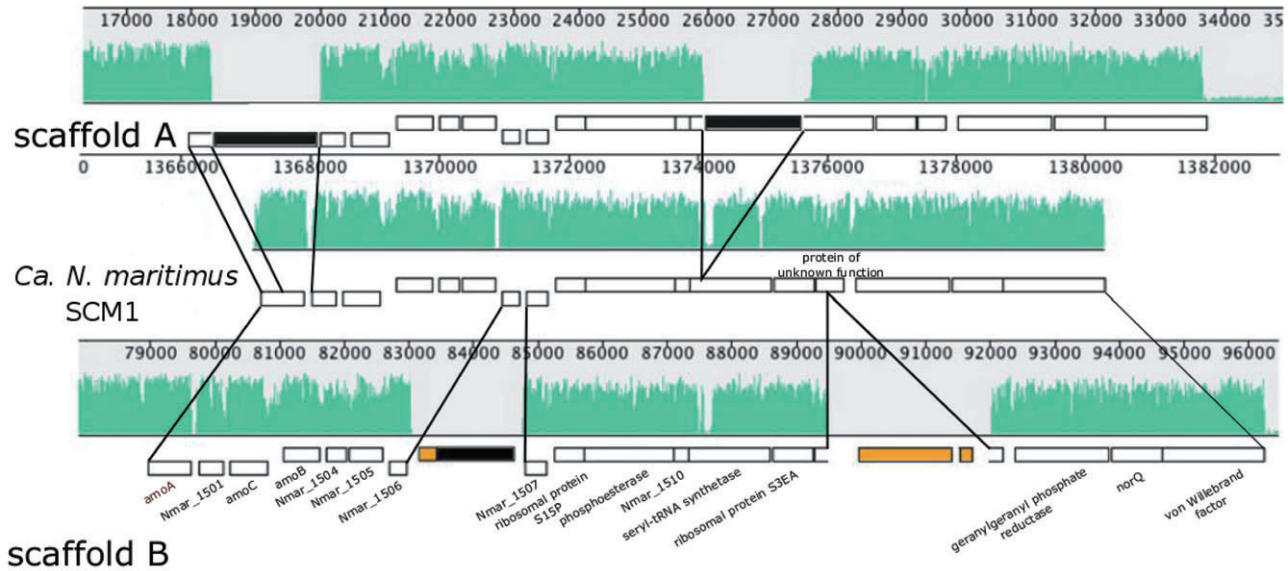

Fig. 6. MAUVE alignments of environmental 'Scaffold A' and 'Scaffold B' compared with '*Ca.* N. maritimus SCM1'. White bars indicate annotated genes. Black bars represent internal sequencing gaps. Orange bars indicate genes unique to the scaffold. Colours on horizontal axis illustrate aligned regions.

amoB and *amoC* on 'Scaffold B' had 89.1% and 87.3% NAID to '*Ca.* N. maritimus SCM1'. Despite similar degrees of divergence on both scaffolds for *amoB* and *amoC* when compared with '*Ca.* N. maritimus SCM1', there is still substantial divergence between the two scaffolds for both genes, 97.3% and 95.4% NAID respectively. This indicates that the differences on each scaffold are variations unique to that scaffold (Table S3).

Two scaffolds contained the region with ^{4}OH-butyryl-CoA dehydratase. 'Scaffold C' (17 950 bp) and 'Scaffold D' (166 197 bp) are syntenic, yet contain an inversion relative to the '*Ca.* N. maritimus SCM1' genome, and possess a non-protein-coding region (Fig. 7). Despite this high degree of synteny, the two scaffolds are divergent to each, in a similar fashion as the genes on 'Scaffold A' and 'Scaffold B'.

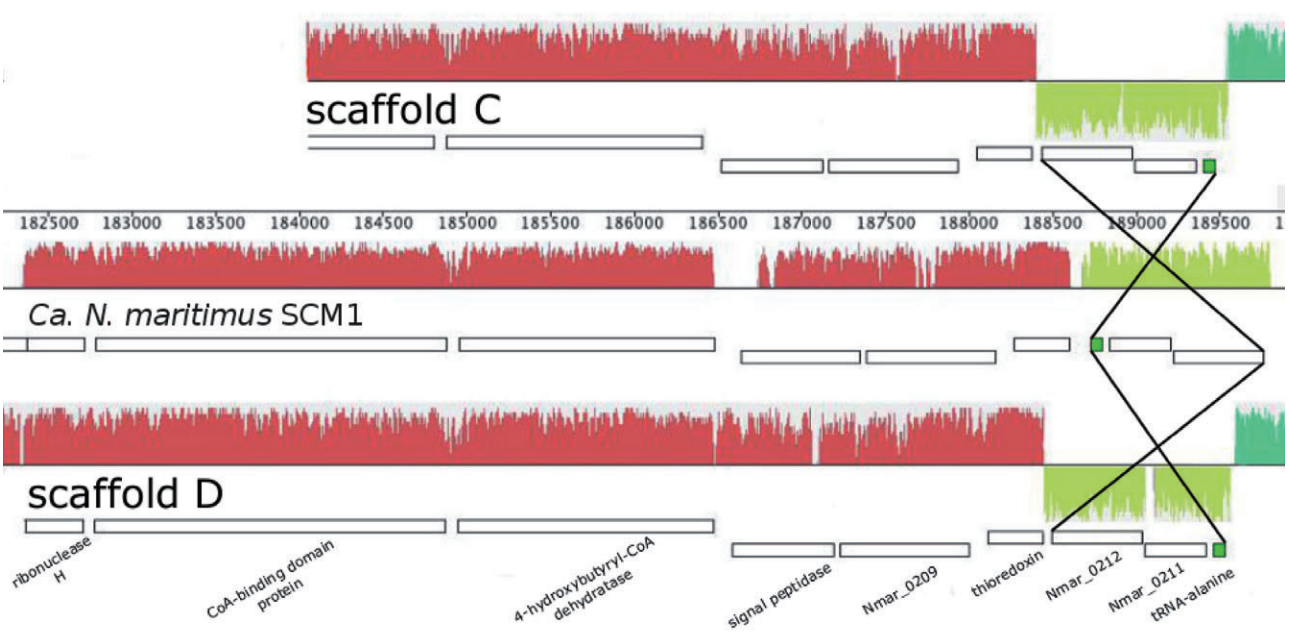

Fig. 7. MAUVE alignments of environmental 'Scaffold C' and 'Scaffold D' compared with '*Ca.* N. maritimus SCM1'. White bars indicate annotated genes. Green bar indicates tRNA sequence. Colours on horizontal axis illustrate aligned regions.

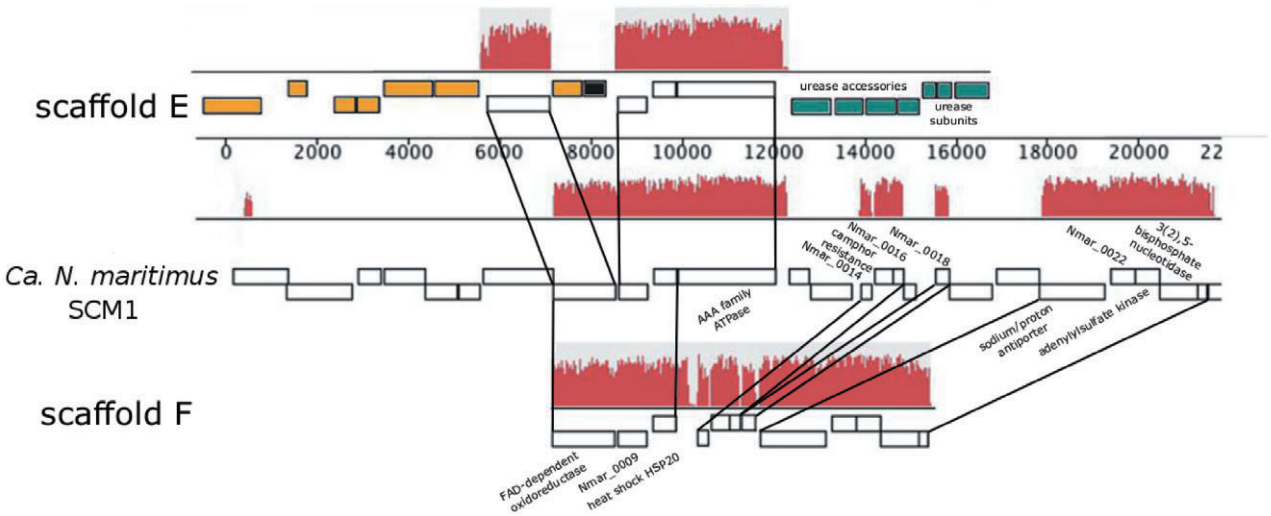

Fig. 8. MAUVE alignments of environmental 'Scaffold E' and 'Scaffold F' compared with '*Ca.* N. maritimus SCM1'. White bars indicate annotated genes. Black bars represent internal sequencing gaps. Orange bars indicate genes unique to the scaffold. Turquoise bars indicate genes in the urease operon. Colours on the horizontal axis illustrate aligned regions.

One scaffold was identified to contain the full urease operon. 'Scaffold E' (16 428 bp) contains the urease operon (urease subunits α, β, γ, and accessory proteins E–H) and has the highest nucleotide similarity to '*Ca.* C. symbiosum A'. 'Scaffold E' had the least overall synteny to the '*Ca.* N. maritimus SCM1' genome, with only four syntenic genes (Fig. 8). The other genes present on the scaffold had 31–71% AAID to genes identified in '*Ca.* N. maritimus SCM1', '*Ca.* C. symbiosum A', and a fosmid clone derived from an uncultivated deep-sea *Thaumarchaea* (HF4000_APKG8I13) (Konstantinidis and DeLong, 2008), supporting the thaumarchaeal origin of the scaffold. The syntenic genes on 'Scaffold E' were present on another environmental scaffold. 'Scaffold F' (41 439 bp) maintains the synteny of the '*Ca.* N. maritimus SCM1' genes for this region, but has several deletions, and has no synteny over the rest of the scaffold. The pattern of gene similarity seen between '*Ca.* N. maritimus SCM1' and the other environmental scaffolds is present for the three genes that 'Scaffold E' and 'Scaffold F' have in common (Table S3).

Collectively, the analysis of these three different regions, and six different scaffolds, show that the putative dominant thaumarchaeal populations in the GOM are divergent from each other, just as the ML tree and 16S rRNA sequences suggested. This divergence manifests in two different ways: (i) sequence divergence and (ii) functional divergence. In terms of sequence divergence, for each of the genes analysed, divergence between the scaffolds and '*Ca.* N. maritimus SCM1' were nearly constant, but the sequences on the scaffolds had moderate nucleotide divergence from each other, suggesting that the two populations are drifting in different directions, in relation to '*Ca.* N. maritimus SCM1'. Results from δN/δS calculations, indicate that all of the analysed genes were undergoing purifying selection, this suggests that the functionality of the genes may be remaining constant. Furthermore, only 'Scaffold C' and 'Scaffold D' were syntenic for sequence order. The four other environmental scaffolds had several insertions and deletions relative to each other, suggesting that these populations are becoming more divergent through environmental gene loss and gain. The scaffolds illustrate one clearly defined example of functional divergence ('Scaffold B' may be functionally different from 'Scaffold A', depending on the activity of the two hypothetical proteins) that 'Scaffold E' has the genomic potential to utilize urease in the environment as a nitrogen source, while 'Scaffold F' does not.

The population diversity that is part of the GOM thaumarchaeal community is not unlike the diversity seen in other globally distributed planktonic marine prokaryotes. The high 16S rRNA gene sequence NAID, indicating a single species, masks a great deal of underlying complexity. The analysis provided suggests that the putative dominant populations in the environment are closely related to each other and '*Ca.* N. maritimus SCM1'. The scaffolds show that much of the underlying gene content remains constant for the dominant environmental populations, and identical features, such as rearrangements and insertions, are present in both. Yet, despite this high degree of similarity on the gene content level and evolutionary selection pressures, the environmental populations are clearly unique, each possessing divergent nucleotide identity from '*Ca.* N. maritimus SCM1' and from each

other. The presence of the full urease operon likely results in a distinct ecological niche for the population possessing this island, a niche that allows for divergent evolution between closely related populations.

Experimental procedures

Sample collection

Sea surface water samples were collected from the GOM at three sites (GOM03, GOM04 and GOM06) in January of 2006 from the R/V Delaware and three sites (GOM12, GOM13 and GOM14) in August of 2006 from the R/V Albatross IV (Table 1). Samples were collected using the JCVI standard operating procedure (Rusch *et al*., 2007). Briefly, 200 l of surface water, from approximately 1.5 m depth, was passed through a 25 μm Nytex pre-filter. The sample was then size-fractionated by filtering sequentially through 3.0-μm-, 0.8-μm- and 0.1-μm-pore-size filters (Supor membrane disc filter, Pall Life Sciences). Filters used for genomic extractions were placed in buffer and immediately frozen in liquid N_2 on the vessel, and transferred to –80°C freezer until DNA isolation could be performed.

Sequencing and assembly

Sample processing proceeded as described in Rusch and colleagues (2007). In brief, DNA was collected via a freeze–thaw method in an EDTA lysis buffer followed by a phenol/chloroform extraction from the 0.1 μm filter, fragmented via nebulization, ligated to BstXI adapters, inserted in to BstXI-linearized medium copy pBR322 plasmids vectors with a medium range insert size, and electroporated into *Escherichia coli*. Following cloning, single colonies were grown overnight in 2 ml of liquid media, lysed by an alkaline lysis miniprep, and DNA was collected by isopropanol precipitation. Paired-end Sanger sequencing was performed from the plasmids using standard M13 forward and reverse primers. A total of 2 827 702 reads were returned (453 807 from GOM03, 957 738 from GOM04, 10 041 from GOM06, 470 592 from GOM12, 925 795 from GOM13 and 9729 from GOM14) containing over 2235 Mbp of sequence data. To reduce redundancy, the sequences were assembled with the Celera assembler at the J. Craig Venter Institute as described by Rusch and colleagues (2007). In total, about 15.5% of sequences assembled in scaffolds greater than 10 kb.

Assembly of the thaumarchaeal bin (see below) was performed using the Celera assembler using a specification file designed to construct microbial-sized genomes from metagenomic data. The following settings were changed:

utgErrorRate = 0.08
ovlErrorRate = 0.10
cnsErrorRate = 0.10
cgwErrorRate = 0.10
merSize = 14
utgGenomeSize = 2 000 000

Determination of the total length spanned by assembled scaffolds is possible due to approximations made of internal sequencing gaps generated using paired-end reads and the Celera assembler.

Identification of thaumarchaeal signature

The GOM metagenome scaffolds and degenerate contigs were searched (BLASTN; Altschul *et al*., 1997) against the '*Ca.* N. maritimus SCM1' genome (1 645 259 bp, 1795 putative CDS, Accession No. NC_010085) using varying levels of stringency. All scaffolds containing any alignment region with an *E*-value $\leq 10^{-3}$ (929 scaffolds totalling 9 208 528 bp) were submitted to the J. Craig Venter Institute Annotation Service and produced 10 990 CDS. All CDSs returned through the annotation pipeline were searched (BLASTP; *E*-value cut-off $\leq 10^{-10}$) against the NCBI nr protein database and assigned a putative taxonomic and gene assignment based on the highest scoring pair with informative annotation. A two-step screening process refined the bin assignments. The first step removed all scaffolds where less than 50% of the CDS had a best BLASTP hit (based on bit score) to a sequenced marine thaumarchaeal genome or fosmid, including '*Ca.* N. maritimus SCM1', '*Ca.* C. symbiosum A', uncultured *Crenarchaeota* 74A7 (Béjá *et al*., 2002), 4B7 (Béjá *et al*., 2002), DeepAnt-EC39 (Lopez-Garcia *et al*., 2004) and HF4000 (Konstantinidis and DeLong, 2008). This initial screen yielded 742 scaffolds containing > 2.4 Mbp. A final quality control step temporarily removed from consideration all CDS which had < 50% identity to sequenced marine thaumarchaeal genomes or fosmids across the length of the BLASTP alignment, and then scaffolds were then reassessed using the first-step curation criteria. This step filtered out four additional scaffolds that contained only sequences with poor alignments to known thaumarchaeal sequences. The small size (e.g. < 3 putative CDSs) and poor alignments of these scaffolds suggested high divergence from the marine *Thaumarchaeota*. The initial planktonic thaumarchaeal bin contained 1324 of 1795 genes (73.7%) annotated in the '*Ca.* N. maritimus SCM1' genome. To assess the quality of our binning strategy, TETRA (Teeling *et al*., 2004) was used to compare the tetranucleotide frequency z-scores between the binned scaffolds and the '*Ca.* N. maritimus SCM1' genome.

The assembled scaffolds likely did not contain all the possible thaumarchaeal diversity. Metagenomic assemblies generated using the Celera Assembler can result in the most abundant organisms being assigned a degenerate contig flag. Therefore, the degenerate contigs of the initial GOM metagenomic assembly were of interest due to the unique population structure of the planktonic thaumarchaeal scaffolds.

The resulting *Thaumarchaea*-like scaffolds (738) and contigs (3662) (36 111 sequences totalling 30 258 762 bp) were recruited with the Geneious (V.4.8.3) (Drummond *et al*., 2009) assembly program with the High Sensitivity parameter and using the '*Ca.* N. maritimus SCM1' as a reference genome. In brief, the Geneious assembly algorithm determines all pairwise distance in a BLAST-like search and progressively aligns the highest scoring pairs. The High Sensitivity parameter increases the time necessary to perform the assembly, but results in a more accurate alignment of sequences to each other after initial alignment to the reference sequence.

Fragment recruitment plots, as described in Rusch and colleagues (2007), were generated comparing all reads in the GOM metagenome and all publicly available GOS sequences

against the '*Ca.* N. maritimus SCM1' genome. In brief, all sequences are BLASTN compared against the reference genome, such that all sequences with $\geq 55\%$ nucleotide identity are displayed along the length of the genome and colour coded to represent specific sampling sites or mate-pair relationships (Rusch *et al.*, 2007). Fragment recruitment plots were assessed for regions divergent from the reference genome.

Functional gene assignment

Amino acid sequences of proteins for the 3-hydroxypropionate/4-hydroxybutyrate cycle from *M. sedula* DSM 5348 (Berg *et al.*, 2007) were obtained from Integrated Microbial Genomes (IMG) (Markowiz *et al.*, 2006) and searched (BLASTX) against all thaumarchaeal reads (*E*-value cut-off = 10^{-3}) and (BLASTP) against the '*Ca.* C. symbiosum A' and '*Ca.* N. maritimus SCM1' genomes (*E*-value cut-off = 10^{-2}). All reads with similarity were considered as putative matches for further consideration. Amino acid sequences of proteins of the ammonia oxidation pathway described in '*Ca.* C. symbiosum A' (Hallam *et al.*, 2006a) were obtained from IMG (Markowiz *et al.*, 2006) and NCBI and searched (BLASTX) against all thaumarchaeal reads. All read matches with an *E*-value $\leq 10^{-3}$ were considered as putative matches for further consideration. Reads identified as possible matches were assessed by comparing the location of identity between the read and amino acid sequences. If a disagreement in the location of the identity was identified (i.e. identity in the middle of the read was matched to identity in the middle of the amino acid sequences, with no homology to the end of the read), reads were no longer considered as a putative match.

Within-bin diversity

To analyse the diversity within the planktonic *Thaumarchaea*, key genes in DNA recombination and repair, aerobic ammonia oxidation, and the ^{3}OH-propionate/^{4}OH-buytyrate cycle (Huber *et al.*, 2008) were compared (BLASTX) against all thaumarchaeal-like reads. Alignments for each gene were generated using CLUSTALW (Thompson *et al.*, 1994) (Cost matrix: IUB; Gap open cost: 16; Gap extend cost: 6.66) and trimmed with Geneious (Drummond *et al.*, 2009) to maximize the length of the overlapping region and the number of sequences included in the alignment. ML trees of sequences with homology to *radA* (21 reads; 476 bp region), *amoA* (16 reads; 417 bp region), methylmalonyl-CoA mutase (15 reads; 394 bp region) and ^{4}OH-butyryl-CoA dehydratase (20 reads; 527 bp region) were constructed using PHYML (Guindon and Gascuel, 2003) [Kimura (K80) model and all other default settings]. Representative environmental sequences were gathered from online databases; the NCBI nt database was used for ammonia monooxygenase and the CAMERA database (V.1.3.2.31; http://camera.calit2.net/) was used for methylmalonyl-CoA mutase and ^{4}OH-butyryl-CoA dehydratase.

Coverage

The coverage depth of each base pair on 738 thaumarchaeal scaffolds was determined using the sum of the length of each read used to construct a scaffold divided by the length of scaffold consensus sequence (read to scaffold coverage). The coverage depth of each base pair in the '*Ca.* N. maritimus SCM1' genome was obtained from the results of the Geneious (Drummond *et al.*, 2009) assembly program.

Thaumarchaeal scaffold analysis

Scaffolds constructed using the thaumarchaeal sequence bin were aligned against the '*Ca.* N. maritimus SCM1' genome using the program MAUVE, and the progressiveMauve alignment algorithm (Darling *et al.*, 2010). Scaffold sequences with homology to coding regions of '*Ca.* N. maritimus SCM1' were compared for pairwise identity in Geneious (Drummond *et al.*, 2009), translated, and assayed for codon alignment (Suyama *et al.*, 2006), which would also calculate synonymous (d_S) and non-synonymous (d_N) substitution rates for the genes using the codeml program PAML (Yang, 2007).

Availability of sequences

The complete GOM metagenome has been deposited in to NCBI GenBank as raw reads in the Trace Archive (TA) (ID No. 2307942905–2310786347). The scaffolds generated from the putative thaumarchaeal sequence bin have been deposited at DDBJ/EMBL/GenBank as a Whole Genome Shotgun project, under the accession AGBE00000000. The version described in this paper is the first version, AGBE01000000.

Acknowledgements

This work was supported by the National Science Foundation Microbial Sequencing Grant 0412119. The authors gratefully acknowledge NOAA ecosystem process division scientists Jon Hare and Jerry Prezario for ship time on NOAA Fisheries R/V Delaware II (Cruise No. DE 06-02) and R/V Albatross IV (Cruise No. AL 06-07). We thank Drs Karla Heidelberg and Shannon Williamson for collecting samples. We thank Robert Friedman, and Yu-Hui Rogers for technical and scientific support in the sequencing efforts. We thank Matt Lewis and Dr Aaron Halpern, who processed the GOM samples. We would like to thank JCVI for providing the JCVI Annotation Service, which provided us with automatic annotation data and the manual annotation tool Manatee.

References

Alber, B.E., Kung, J.W., and Fuchs, G. (2008) 3-Hydroxypropionyl-coenzyme A synthetase from *Metallosphaera sedula*, an enzyme involved in autotrophic CO_2 fixation. *J Bacteriol* **190:** 1383–1389.

Allen, E.E., Tyson, G.W., Whitaker, R., Detter, J.C., Richardson, P.M., and Banfield, J.F. (2007) Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci USA* **104:** 1883–1888.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped BLAST and PSI-

BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Béjá, O., Suzuki, M., Heidelberg, J., Nelson, W., Preston, C., Hamada, T., *et al*. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415:** 630–633.

Berg, I.A., Kockelkorn, D., Buckel, W., and Fuchs, G. (2007) A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. *Science* **318:** 1782–1786.

Berg, I.A., Ramos-Vera, W.H., Petri, A., and Huber, H. (2010) Study of the distribution of autotrophic CO_2 fixation cycles in Crenarchaeota. *Microbiology* **156:** 256–269.

Bernhard, A., Landry, Z., Blevins, A., De La Torre, J., Giblin, A., and Stahl, D. (2010) Abundance of ammonia-oxidizing archaea and bacteria along an estuarine salinity gradient in relation to potential nitrification rates. *Appl Environ Microbiol* **76:** 1285–1289.

Bhaya, D., Grossman, A., Steunou, A.-S., Khuri, N., Cohan, F., Hamamura, N., *et al*. (2007) Population level functional diversity in the microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* **1:** 703–713.

Blainey, P., Mosier, A., Potanina, A., Francis, C., and Quake, S. (2011) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS ONE* **6:** e16626.

Christensen, J., Townsend, D.W., and Montoya, J. (1996) Water column nutrients and sedimentary denitrification in the Gulf of Maine. *Cont Shelf Res* **16:** 489–515.

Christman, G.D., Cottrell, M.T., Popp, B.N., Gier, E., and Kirchman, D.L. (2011) Abundance, diversity, and activity of ammonia-oxidizing prokaryotes in the coastal arctic ocean in summer and winter. *Appl Environ Microbiol* **77:** 2026–2034.

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.

Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5:** e11147.

Drummond, A., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., *et al*. (2009) Geneious v4.6 [WWW document]. URL: http://www.geneious.com/.

Dyhrman, S.T., and Anderson, D.M. (2003) Urease activity in cultures and field populations of the toxic dinoflagellate Alexandrium. *Limnol Oceanogr* **48:** 647–655.

Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8:** 186–194.

Fouts, D.E., Tyler, H.L., DeBoy, R.T., Daugherty, S., Ren, Q., Badger, J.H., *et al*. (2008) Complete genome sequence of the N_2-fixing broad host range endophyte *Klebsiella pneumoniae* 342 and virulence predictions verified in mice. *PLoS Genet* **4:** e1000141.

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 Natl Acad Sci USA* **102:** 14683–14688.

Guindon, S., and Gascuel, O. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52:** 696–704.

Hagström, Ä., Pinhassi, J., and Zweifel, U.L. (2000) Biogeographical diversity among marine bacterioplankton. *Aquat Microb Ecol* **21:** 231–244.

Hallam, S.J., Mincer, T.J., Schleper, C., Preston, C.M., Roberts, K., Richardson, P.M., and DeLong, E.F. (2006a) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* **4:** 520–536.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y., Sugahara, J., *et al*. (2006b) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* **103:** 18296–18301.

Hatzenpichler, R., Lebedeva, E.V., Spieck, E., Stoecker, K., Richter, A., Daims, H., and Wagner, M. (2008) A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring. *Proc Natl Acad Sci USA* **105:** 2134–2139.

Huber, H., Gallenberger, M., Jahn, U., Eylert, E., Berg, I.A., Kockelkorn, D., *et al*. (2008) A dicarboxylate/4-hydroxybutyrate autotrophic carbon assimilation cycle in the hyperthermophilic Archaeum *Ignicoccus hospitalis*. *Proc Natl Acad Sci USA* **105:** 7851–7856.

Ingalls, A.E., Shah, S.R., Hansman, R.L., Aluwihare, L.I., Santos, G.M., Druffel, E.R.M., and Pearson, A. (2006) Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* **103:** 6442–6447.

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

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

Labrenz, M., Sintes, E., Toetzke, F., Zumsteg, A., Herndl, G.J., Seidler, M., *et al*. (2010) Relevance of a crenarchaeotal subcluster related to *Candidatus Nitrosopumilus maritimus* to ammonia oxidation in the suboxic zone of the central Baltic Sea. *ISME J* **4:** 1496–1508.

Lopez-Garcia, P., Brochier, C., Moreira, D., and Rodriguez-Valera, F. (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeota reveals multiple horizontal gene transfers. *Environ Microbiol* **6:** 19–34.

Markowiz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., *et al*. (2006) The integrated microbial genome (IMG) system. *Nucleic Acids Res* **34:** D344–D348.

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

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

Moraru, C., Lam, P., Fuchs, B.M., Kuypers, M.M.M., and Amann, R. (2010) Gene-FISH – an in situ technique for linking gene presence and cell identity in environmental microorganisms. *Environ Microbiol* **12:** 3057–3073.

Myers, E.W., Sutton, G., Delcher, A.L., Dew, I.M., and Fasulo, D.P. (2000) A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Park, H., Wells, G.F., Bae, H., Criddle, C.S., and Francis, C.A. (2006) Occurrence of ammonia-oxidizing archaea in wastewater treatment plant bioreactors. *Appl Environ Microbiol* **72:** 5643–5647.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al*. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5:** 398–431.

Santoro, A., Casciotti, K., and Francis, C. (2010) Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. *Environ Microbiol* **12:** 1989–2006.

Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609–W612.

Teeling, H., Meyedierks, A., Bauer, M., Amann, R., and Glockner, F.O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6:** 938–947.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673–4680.

de la Torre, J., Walker, C., Ingalls, A., Könneke, M., and Stahl, D. (2008) Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* **10:** 810–818.

Tourna, M., Stieglmeier, M., Spang, A., Konneke, M., Schintlmeister, A., Urich, T., *et al*. (2011) *Nitrososphaera viennensis*, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* **108:** 8420–8425.

Van Mooy, B.A.S., Rocap, G., Fredricks, H.F., Evans, C.T., and Devol, A.H. (2006) Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proc Natl Acad Sci USA* **103:** 8607–8612.

Venter, J.C., Remington, K., Heidelberg, J., Halpern, A.L., Rusch, D., Eisen, J., *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

Walker, C.B., de la Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, N., Arp, D.J., *et al*. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* **107:** 8818–8823.

Wuchter, C., Schouten, S., Boschker, H., and Damste, J.S. (2003) Bicarbonate uptake by marine Crenarchaeota. *FEMS Microbiol Lett* **219:** 203–207.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Maximum likelihood tree (bootstrap: 100) of a 403 bp region of the 4-OH-butyryl-CoA gene from 19 GOM sequences (ID starting 110-) and 2 environmental sequences obtained from the CAMERA database (GOS ID#).

Fig. S2. Maximum likelihood tree (bootstrap: 100) of a 419 bp region of the *radA* gene from 21 GOM sequences (ID starting 110-).

Table S1. Sequence recruitment for genes related to the carbon fixation pathway.

Table S2. Location of gaps after reads are recruited to *Ca. N. maritimus* SCM1 genome.

Table S3. Comparison of genes of interest between assembled scaffolds and *Ca. N. maritimus* SCM1.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.