OPEN

ORIGINAL ARTICLE

Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses

Jun Meng[1,3], Jun Xu[1], Dan Qin[1], Ying He[1], Xiang Xiao[1] and Fengping Wang[1,2]

[1]State Key Laboratory of Microbial Metabolism and State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai, People's Republic of China and [2]Key Laboratory of Systems Biomedicine, Ministry of Education, Shanghai Jiao Tong University, Shanghai, People's Republic of China

The Miscellaneous Crenarchaeota group (MCG) Archaea is one of the predominant archaeal groups in anoxic environments and may have significant roles in the global biogeochemical cycles. However, no isolate of MCG has been cultivated or characterized to date. In this study, we investigated the genetic organization, ecophysiological properties and evolutionary relationships of MCG archaea with other archaeal members using metagenome information and the result of gene expression experiments. A comparison of the gene organizations and similarities around the 16S rRNA genes from all available MCG fosmid and cosmid clones revealed no significant synteny among genomic fragments, demonstrating that there are large genetic variations within members of the MCG. Phylogenetic analyses of large-subunit + small-subunit rRNA, concatenated ribosomal protein genes and topoisomerases IB gene (TopoIB) all demonstrate that MCG constituted a sister lineage to the newly proposed archaeal phylum Aigarchaeota and Thaumarchaeota. Genes involved in protocatechuate degradation and chemotaxis were found in a MCG fosmid 75G8 genome fragment, suggesting that this MCG member may have a role in the degradation of aromatic compounds. Moreover, the expression of a putative 4-carboxymuconolactone decarboxylase was observed when the sediment was supplemented with protocatechuate, further supporting the hypothesis that this MCG member degrades aromatic compounds.

Introduction

The study of the ecophysiology of archaea is currently one of the most exciting research areas in the field of environmental microbiology. Many uncultivated archaeal groups have been discovered as microbial diversity surveys have expanded and improved, but the physiological properties of most of these uncultivated archaea remain to be determined. For instance, uncultivated archaeal lineages such as Marine benthic group B (also known as deep-sea archaeal group), Miscellaneous Crenarchaeota group (MCG) or South African gold mine Euryarchaeota group were found widespread in marine sediments (Teske and Sørensen, 2008); however, the functions of these archaea in the environments are still unknown.

MCG archaea live in diverse habitats, including terrestrial and marine, hot and cold, surface and subsurface environments (Biddle et al., 2006; Teske, 2006; Kubo et al., 2012). The label 'miscellaneous' appears to represent the difficulty in categorizing the wide terrestrial and marine habitat range of this group (Inagaki et al., 2003). Sørensen and Teske (2006) divided hundreds of MCG clones into smaller and more manageable subgroups—MCG-1 to MCG-4. Jiang and Li, (2011) performed a comprehensive phylogenetic analysis and divided MCG archaea into seven subgroups (MCG-A to MCG-G), whereas Kubo et al. (2012) divided MCG archaea into 17 subgroups. In addition to its cosmopolitan distribution, the MCG group of archaea is one of the most abundant groups in the subsurface sedimentary biosphere based on the 16S rRNA gene abundance: the MCG clones account for 33% of all clones from 47 16S rRNA gene libraries obtained from 11 published studies of the deep marine biosphere (Fry et al., 2008). Moreover, the MCG was found to be one of the most active groups in the deep marine biosphere (Fry et al., 2008; Li et al., 2012b). Parallel 16S rRNA and rDNA analyses of Ocean Drilling Program site 1229 on the Peru Margin indicated that the MCG dominated the archaeal clone libraries based on PCR-amplified 16S rDNA genes

(Parkes et al., 2005) and on reverse-transcribed 16S rRNA (Biddle et al., 2006). At the Ocean Drilling Program site 1227 on the Peru Margin, MCG archaea were abundant in 16S rRNA gene clone libraries from all depths (Inagaki et al., 2006) and they dominated the reverse-transcribed 16S rRNA pool in all sediment layers except the deep-sea archaeal group/Marine benthic group B horizon (Sørensen and Teske, 2006). In addition, the carbon isotope signatures of archaeal cells and polar lipids from MCG-dominated sediment horizons indicate that these anaerobes utilize buried organic carbon substrates (Biddle et al., 2006). The widespread distribution, high abundance and metabolic activities of MCG archaea all indicate that these organisms might be significant players in biogeochemical cycles. However, the paucity of representative pure cultures has hindered our understanding of the physiological properties of these archaea as well as their ecological functions and evolutionary position. Environmental genomics provides an approach to explore the potential physiological characteristics and genomic information of uncultivated microbes in the context of indigenous microbial communities. Just recently, single-cell genome analysis suggested that members of the MCG archaea are specializing in extracellular protein degradation (Lloyd et al., 2013). Till now, only a few MCG fosmid and cosmid clones have been identified. One MCG fosmid clone was reported containing a functional bacteriochlorophyll a synthase (bchG) gene, a key enzyme for bacteriochlorophyll a biosynthesis. However, the in vivo physiological functions of BchG in MCG are still unknown, although it was supposed that containing a presumptive Bchl a synthase gene, may give the archaea more flexibility to survive or adapt to various environments (Meng et al., 2009). The other three analyzed fosmid clones contain homologous to potentially important functional genes involving in lipid biosynthesis, energy metabolism and resistance to oxidants (Li et al., 2012a). But the physiological properties and the roles of these organisms in natural biogeochemical cycles are still remaining to be determined.

In this study, we investigated the phylogenetic position and potential ecophysiological properties of this little understood MCG archaeal group using an environmental metagenomic method. A member of the MCG was hypothesized to be aromatic compound degrader based on genome information. This hypothesis was further supported by target gene expression analysis after substrate supplementation.

Materials and methods

Site description and sampling
Estuarine sediment was collected from a site of around 0.5 m water depth on the Qi'ao Island (Pearl River Estuary, 22°27′21.4″ N, 113°38′7.3″E) in Guangdong Province, China, in 2005 April using a single-core sampler. The temperature of the bottom water was 21.5 °C and the salinity at the surface of sediment was 2.6%. Mangrove sediment was collected from a national mangrove reserve in Jiulongjiang estuary (24°24′48.6″N, 117°56′30.5″E), Fujian Province, China, 2009. All samples were kept on dry ice during transport and then stored in a − 20 °C fridge.

Construction of the genomic library
High-molecular-weight genomic DNA was extracted according to the method of Zhou et al. (1996) and separated using pulsed-field agarose gel electrophoresis after both DNA ends were end-repaired following the manufacturer's instructions (Epicentre, Madison, WI, USA). After the electrophoresis was completed, an agarose plug containing 33–48 kb DNA was cut out, and the DNA was recovered using electro-elution (Bio-Rad, Hercules, CA, USA). The genomic DNA purified from this plug was ligated to pCC1FOS fosmid or pWEB-TNC cosmid, followed by packaging into MaxPlax Lambda Packaging Extract (Epicentre). The packaged particles were transferred into Escherichia coli EPI300 or EPI100 (Epicentre). In total, ∼ 8000 clones for the estuarine sediment and ∼ 9000 clones for the mangrove sediment were obtained in this study. The average insert size was 35 kb.

Screening for the archaeal genome fragments
The library was pooled into groups of 12 clones, and the mixed fosmid or cosmid plasmids were extracted using a standard alkaline lyses procedure. These extracted plasmids were used as templates for PCR amplification. Multiplex PCR with archeal 16S rRNA universal primer set Arch21F/958R (DeLong, 1992) was used to screen for clones containing archaeal 16S rRNA gene. Plasmids of 12 individual fosmid/cosmid clones, with positive archaeal 16S rRNA gene amplification, were then extracted and used as templates for the second round of PCR amplification. The single fosmid/cosmid clones containing archaral 16S rRNA gene were under subsequent investigations.

Analysis of the metagenome sequences 75G8 and 26B6: tRNA genes, Open Reading Frame search and protein identification
Shotgun libraries were sequenced by the Sanger sequencing method to determine the complete insert sequences of each clone as described before (Meng et al., 2009). Open Reading Frame was predicted with GeneMark (Lukashin and Borodovsky, 1998). BLAST were used to search for similar sequences in GenBank (Altschul et al., 1997) with an E-value cutoff of $< 10^{-5}$. In addition, protein annotation against Pfam (Sonnhammer et al., 1997) and InterPro (Zdobnov and Apweiler, 2001) was performed with an e-value $< 10^{-5}$. Signal peptides were scanned

652

with SignalP 4.0 (Petersen *et al.*, 2011) and transmembrane segments were predicted using TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/). The conserved domains of predicted protein sequences were detected using the program InterProScan (Zdobnov and Apweiler, 2001). tRNAs were scanned using tRNAscan-SE v.1.21 tool (Lowe and Eddy, 1997).

Phylogenetic analysis of 16S rRNA, LSU-SSU rRNA and predicted proteins
For 16S rRNA phylogeny, representatives of MCG and Marine benthic group B, Marine group I were selected from ARB-silva (http://www.arb-silva.de/) as reference sequences. The LSU-SSU (large-subunit-small-subunit) operon (23S rRNA-16S rRNA) from Crenarchaeota, Euryarchaeota, Thaumarchaeota from GenBank were selected for LSU-SSU phylogenetic tree. MAFFT with L-INS-i strategy was used for all alignments in this paper (Katoh *et al.*, 2002). Maximum likelihood phylogenetic trees of aligned genes were inferred with RAxML, using the general time-reversible model of substitution and the GAMMA model of rate heterogeneity; tree topologies were checked by 100 bootstrapping replicates.

In the case of ribosomal proteins, each ribosomal protein was aligned by MAFFT first and then all alignments were concatenated. For the phylogenetic tree of protein, the best protein model was determined with ProteinModelSelection.pl (http://sco.h-its.org/exelixis/software.html). LG were selected as best protein model for ribosomal protein and Topomerase IB (ToPoIB) protein (Le and Gascuel, 2008). Maximum likelihood phylogenetic trees were constructed using RAxML estimated by the LG model of protein substitution and the GAMMA model of rate heterogeneity, with 100 replications for bootstrapping.

Sediment collection and substrate feeding experiment
Mangrove sediment from the same location as that for cosmid library was utilized for substrate feeding experiment, on 2009 October. First, we emptied the core rod of a sterile 50 ml syringe, and then vertically inserted it into sediment with the tip up. After the syringe was filled with sediment, the core rod was pushed back to expel remaining air in the syringe. The top tip of each syringe was immediately sealed with parafilm and all filled syringes were transported on ice to the laboratory. One syringe was stored at $-80\,^{\circ}\text{C}$ as original sample. Others were processed with incubation experiment. Protocatechuate solution was prepared as following: 0.5 g protocatechuate was dissolved in 2 ml of deionized water and filtered with 0.22 mm filter. The protocatechuate solution was then injected into the syringe from the top tip and seeped into the sediment. The tip of the syringe was then sealed with parafilm, and the whole syringe was covered

with foil and incubated at a thermostatic room $(26\,^{\circ}\text{C})$ for 45 days. One syringe without injection of protocatechuate was used as a control and was covered with foil and incubated under the same condition.

RNA extraction and gene expression
Original sample, control sample and each layer (L1–L4) from the protocatechuate-supplemented samples were used for RNA extraction, respectively. Two grams of each sediment sample were used for RNA extraction using E.Z.N.A. Soil RNA Kit (Omega, Bio-Tek, Norcross, GA, USA). Total RNA was treated with DNase I at $37\,^{\circ}\text{C}$ for 1 h to remove potential DNA contamination. Reverse transcription-PCR was performed on the purified total RNA using RevertAid H Minus Reverse Transcriptase (Fermentas, Hanover, MD, USA) with specific primers (CDS16-forward: 5′-CCTCGGCGAGCATTTCCGGG-3′, CDS16-reverse: 5′-GCCCATCGGCAGGAAGGTGG-3′; CDS17-forward: 5′-CATCACCTGCTTGATGCTCT-3′, CDS17-reverse: 5′-CGGGAAATTCGTGGAATATG-3′), following the manufactures' instructions. Two microliters of reaction mixture from reverse transcription was used for following PCR amplification. The PCR cycle condition was as following: $98\,^{\circ}\text{C}$ for 30 s, 30 cylces of $95\,^{\circ}\text{C}$ for 30 s, $55\,^{\circ}\text{C}$ for 30 s, $72\,^{\circ}\text{C}$ for 30 s and extension at $72\,^{\circ}\text{C}$ for 7 min. After amplification, PCR products were subject to electrophoresis, and PCR bands from agarose gels were purified by E.Z.N.A Gel Extraction kit (Omega, Bio-Tek). The purified PCR amplicons were ligated with pMD-18T vector and transformed to *E. coli* DH- 5α. Three positive clones for each PCR amplicon were sent out for sequencing.

Nucleotide sequence accession number
The 16S rRNA gene and the genomic sequences in this study were all deposited in the DDBJ/EMBL/GenBank nucleotide sequence databases with KF439060 and KF439061.

Results and discussion

*Metagenomic library construction and screening*
A cosmid library was constructed from mangrove sediment from Zhangjiang Mangrove Reservation, Fujian Province, China. The mangrove sediment used in this study contained abundant MCG archaea estimated by 16S rRNA gene library analyses (Zhang *et al.*, 2009; Li *et al.*, 2012b). The cosmid library contained ~9000 clones and the average insert length was 35 kb.

This cosmid library derived from the mangrove sediment and a fosmid library constructed from estuarine sediment (Meng *et al.*, 2009) were screened for MCG clones by PCR amplification. Six clones containing archaeal 16S rRNA genes were obtained, four of them belonged to MCG group

(as shown in Figure 1). Three clones 37F10, 75G8 and 26B6 yielded full insert sequences. On the basis of the classification and phylogenetic analyses of MCG 16S rRNA gene sequences by Jiang *et al.* (2011), 37F10 was grouped into MCG-A, whereas 75G8 and 26B6 were placed within the MCG-G
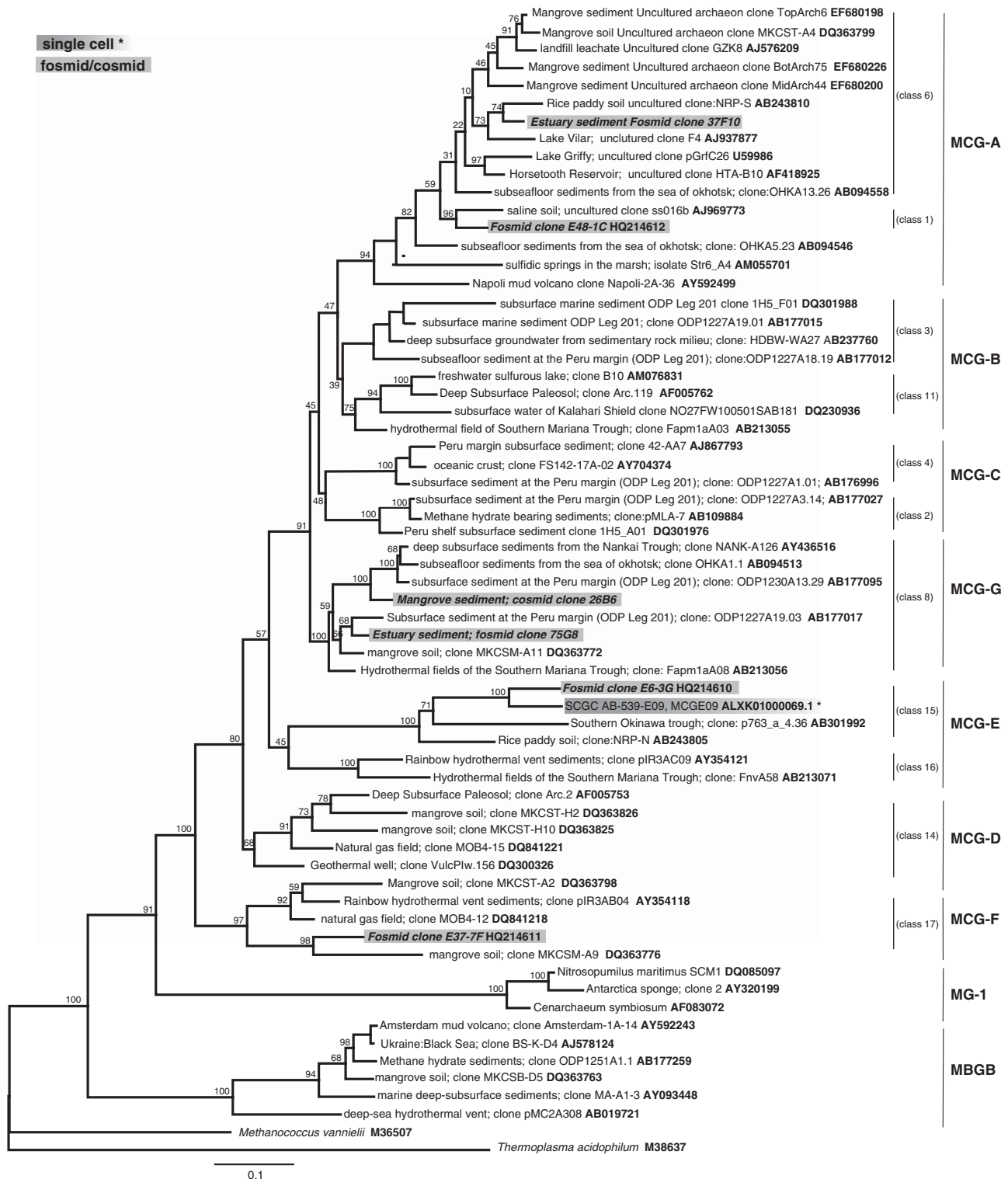
Figure 1 The phylogenetic tree of uncultivated MCG discussed in the text. The tree was constructed from the alignment of >900 unambiguously aligned base pairs using MAFFT followed by Maximum likelihood method by RAxML with the GTRGAMMA model. The stability of the topology was evaluated by bootstrapping (100 replicates). The resulting bootstrap values are indicated at each node in the tree. The names of MCG groups (MCG-A to -G, and class 1–17) were modified based on Jiang *et al.'s* classification (Jiang *et al.*, 2011) and Kubo *et al.*'s classification (Kubo *et al.*, 2012), respectively.

subgroup. Whereas according to Kubo et al. (2012) classification, 37F10 belongs to class 6, 75G8 and 26B6 belong to class 8 (Figure 1).

*Gene composition and comparative analyses of MCG genomic fragments*
Clone 75G8 had a 33887 bp insert size, which contained 32 predicted conserved domain sequences (CDS) and one 16S–23S rRNA operon (Figure 2, Supplementary Table S1). The G+C content of the whole insert was 56.94% and that of 16S rRNA was 59.80%. Twenty-five of the predicted protein-encoding sequences (CDS) could be assigned with functions, four were identified as hypothetical conserved proteins and three of the CDSs did not show significant similarity to any amino-acid sequences in the protein databases.

Clone 26B6 had an insert size of 34887 bp and contained a dispersed 16S rRNA gene and 40 CDSs (Figure 2, Supplementary Table S2). The average G+C content of the insert sequence was 44.71% and that of 16S rRNA gene was 59.15%. Neither 23S nor 5S genes were found in this fragment. Most of the known archaea had one or a few copies of an rRNA operon containing at least both 16S and 23S rRNA genes, but the dispersed localization of the 16S and 23S rRNAs was common within MCG and

other archaeal members (Meng et al., 2009; Li et al., 2012a). Twenty-two of the predicted CDSs could be assigned with functions, 10 were identified as hypothetical conserved proteins and 9 CDSs did not show significant similarity to any amino-acid sequences in the protein databases (Supplementary Table S1).

The main characteristics of six available MCG fosmid/cosmid clones were listed in Table 1. The G+C contents of listed MCG 16S rRNA genes were relatively stable, ranging from 55.9 to 59.8%. But the G+C contents of the whole clones exhibited greater differences to as high as 19.4%, with the lowest G+C content of 37.5% for E37-7F and the highest of 56.9% for 75G8. The similarities between 16S rRNA genes from metagenome clones and single cell clone ranged from 80 to 95% (Table 2), with an average similarity of 85%. The gene organizations on these six fosmid/cosmid MCG fragments were compared (as shown in Figure 2), big variations were demonstrated. Even focusing on the gene organizations around the 16S rRNA gene from all the six MCG metagenome clones, no synteny was found (Figure 2). This result was consistent with previous report that no colinear regions were found between MCG fosmids and any reported archaeal genomic fragments or genomes (Li et al., 2012a).

The physiological traits and ecological significance of MCG archaea remain unclear.
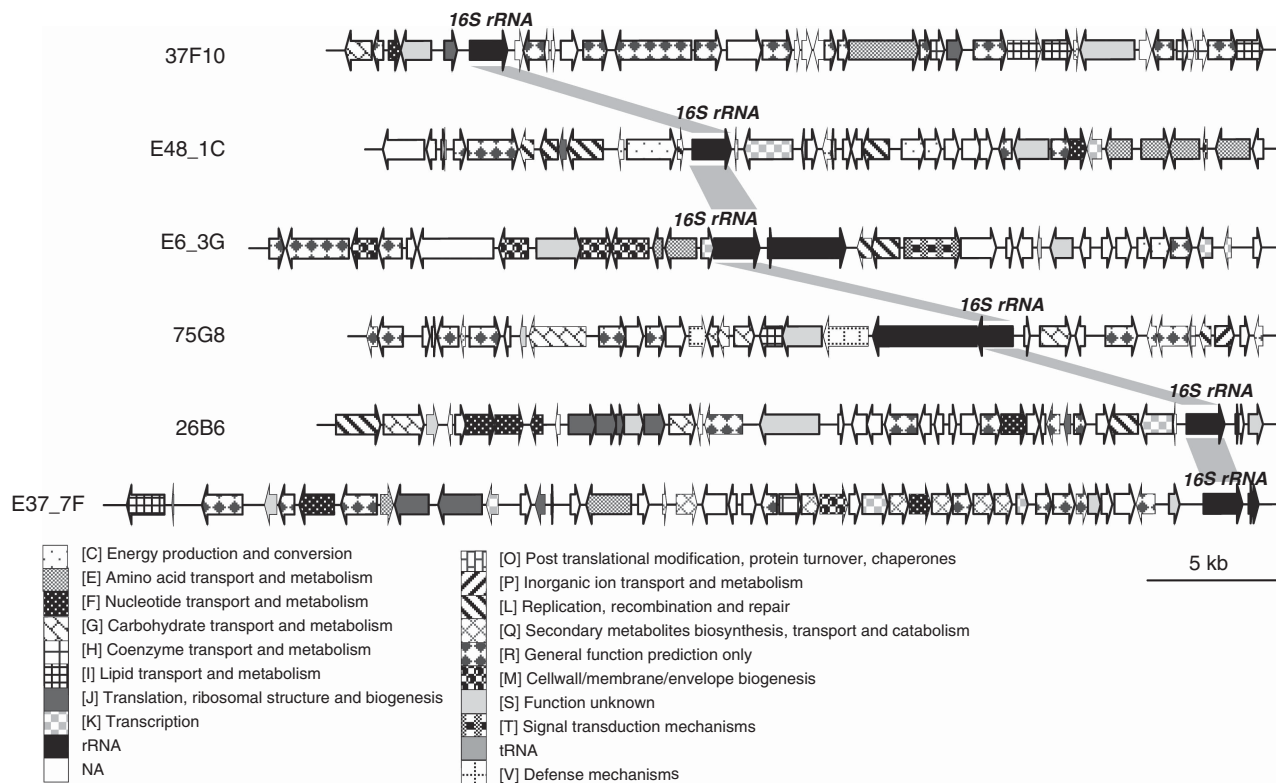
Figure 2 Comparison of gene organization. The gene organizations of the genomic fragment from six MCG fosmid/cosmid clones were compared with each other. The genes are colored according to Clusters of Orthologous Groups (COG) category, and 16S rRNAs are linked in gray.

Table 1 Characteristic summary of MCG fosmid/cosmid clones

	37F10	75G8	26B6	E6-3G	E37-7F	E48-1C
Insert length (bp)	34 528	33 887	34 877	38 227	42 618	34 738
G + C content (%)	52.5	56.9	44.7	50.1	37.5	44.9
G + C content of 16S rRNA (%)	58.0	59.8	59.1	57.6	55.9	58.6
rrn operon and tRNA	16S	16S–23S tRNA^{Ile}	16S tRNA^{Ala}, tRNA^{Trp}	16S-5.8S–23S	16S-5.8S–23S tRNA^{Arg}, tRNA^{Cys}	16S tRNA^{Ser}
No. of predicted ORFs	36	32	40	30	41	37
No. of conserved hypothetic protein	4	4	10	3	4	4
No. of hypothetical protein	7	3	9	12	8	13
Average ORF length	807	756	747	885	696	706
16S rRNA identities to 37F10 (%)	100	87	86	82	82	89

Abbreviations: ORF, Open Reading Frame; MCG, Miscellaneous Crenarchaeota group.

Previous studies have suggested that MCG were distributed in various habitats and exhibit extraordinary versatility. The 16S rRNA sequences of MCG members varied greatly, exhibiting as low as 76% similarity even within groups (Fry *et al.*, 2008; Kubo *et al.*, 2012). The comparison of retrieved MCG genomic fragments indicated huge variations in genomic regions other than the 16S rRNA gene sequences, and such high genomic diversity also supported the high metabolic diversity of MCGs, as suggested by their evolutionary diversity (Biddle *et al.*, 2006; Fry *et al.*, 2008; Teske and Sørensen, 2008).

MCG phylogenetic analysis based on LSU-SSU rRNA, ribosomal proteins and DNA TopoIB gene
Currently, the MCG cluster exhibits no clear affiliation to any of the established archaeal phyla and presented an unstable branching order when 16S rRNA-based trees are constructed with different methods (Pester *et al.*, 2011). LSU-SSU rRNA and/or concatenated ribosomal proteins have served as robust gene markers for phylogenetic analysis. MCG clone 75G8 contained a complete 16S–23S rRNA operon, and clone 26B6 contains several ribosomal protein genes, which gave us a chance to re-examine the phylogenetic relationship of MCG with other archaeal groups.

In the LSU-SSU rRNA phylogenetic tree (Figure 3), MCG was clearly shown as a sister lineage of Aigarchaeota and Thaumarchaoeta. The novel archaeal phylum Aigarchaeota was just recently proposed by Nunoura *et al.* (2011) based on the distinct genomic features (that is, including genes encoding a ubiquitin-like protein modifier system) of Candidatus '*Caldiarchaeum subterraneum*', which belongs to the HWCGI group. Our LSU-SSU rRNA phylogenetic analysis supported the independence of Aigarchaeota from Thaumarchaeota and further suggested that MCG could constitute a new phylum. Moreover, consistent with the LSU-SSU rRNA phylogenetic tree, the phylogenetic tree of concatenated ribosomal proteins (Supplementary Figure S1) also indicated that MCG represents as a sister lineage with Thaumarchaeota.

The genomic fragment of clone 26B6 contained a putative DNA TopoIB type protein (CDS1), which

**Table 2** Similarity of 16S rRNA genes between MCG fosmid/cosmid and single-cell clones

	37F10	75G8	26B6	E6-3G	E37-7F	E48-1C	MCGE09[a]
37F10	1	0.87	0.86	0.82	0.82	0.89	0.80
75G8		1	0.95	0.85	0.87	0.89	0.83
26B6			1	0.85	0.85	0.87	0.83
E6-3G				1	0.83	0.83	0.93
E37-7F					1	0.85	0.81
E48-1C						1	0.83
MCGE09							1

Abbreviation: MCG, Miscellaneous Crenarchaeota group.
[a]MCGE09 is the single-cell sequencing clone (Lloyd *et al.*, 2013), the partial 16S rRNA gene (900 bp) is used here for the pairwise comparison.

showed highest similarity (40%) with that from *Nitrosopumilus maritimus* SCM1 (Supplementary Table S2). Historically, type IB topoisomerases were thought to be eukaryotic-specific enzymes. A shorter version was then found in viruses and later on in several bacteria (Forterre *et al.*, 2007), but these genes were not found in any archaea until recently in members of the proposed novel archaeal phylum Thaumarchaeota and Aigarchaeota (Brochier-Armanet *et al.*, 2008; Nunoura *et al.*, 2011). ToPoIB from MCG clone 26B6 formed a sister group with those from Thaumarchaeota and Aigarchaeota, forming an archaeal branch independent of those from Eukarya and virus (Figure 4). The archaeal-topoisomerase does not seem to be acquired via lateral gene transfer from Eukarya, but instead might have been present in the last ancestral common ancestor.

According to the phylogenetic analyzes of LSU-SSU rRNA, ribosomal proteins and ToPoIB gene, MCG is clearly shown as a sister lineage with Thaumarchaeota and Aigarchaeota, and it is likely that they evolved from a common ancestor. In a recently published partial genome (with 30% genome recovery) obtained from a single cell of a MCG member (MCGE09, as shown in Figure 1), initial phylogenetic analyses using single copy genes in archaea also placed MCG as a sister lineage with Thaumarchaeota and Aigarchaeota (Lloyd *et al.*, 2013). All these
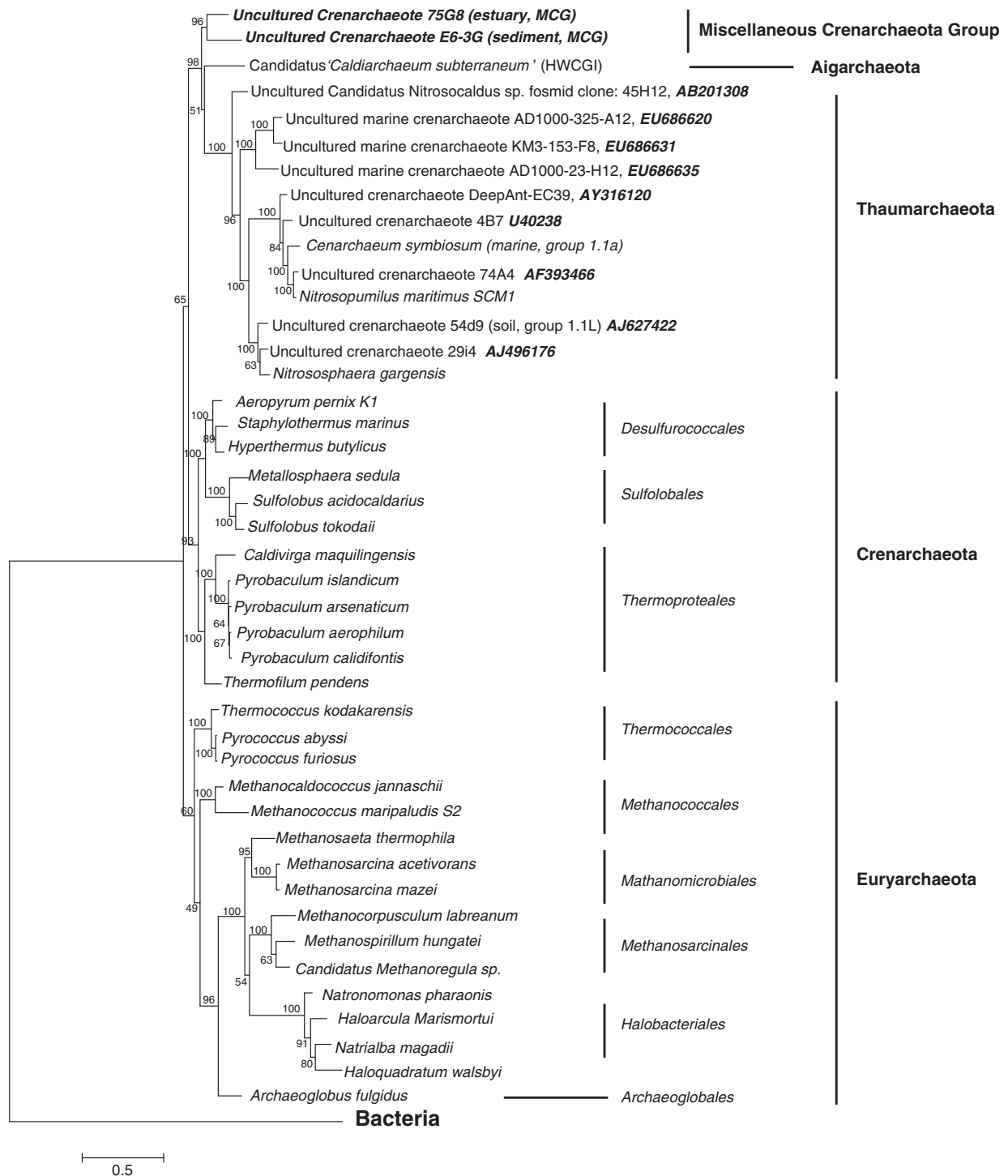
Figure 3 The maximum likelihood tree based on the LSU-SSU sequences from archaea and bacteria. All sequences were retrieved from whole genomes or from environmental genomic fragments that contain LSU-SSU operon. In the tree, 75G8 indicates the genomic fragment obtained in this study. The sequences of LSU-SSU operon were aligned using MAFFT with L-INS-i strategy. The maximum likelihood tree was computed by RAxML program using the general time-reversible (GTR) model of sequence evolution, by including a gamma-correction. The numbers at the nodes represent the non-parametric bootstrap values that were computed by RAxML.

evidences indicate that MCG is not Crenarchaeota, and it locates at a deep branching position with Thaumarchaeota and Aigarchaeaota. Therefore, MCG is likely to be considered as a novel archaeal phylum, and we propose to name the new phylum 'Bathyarchaeota' (from the Greek '*bathys*', meaning deep as it locates deep branching with Thaumarchaeota and Aigarchaeaota, and frequently detected in the deep subsurface sediments). More precise phylogenetic placement
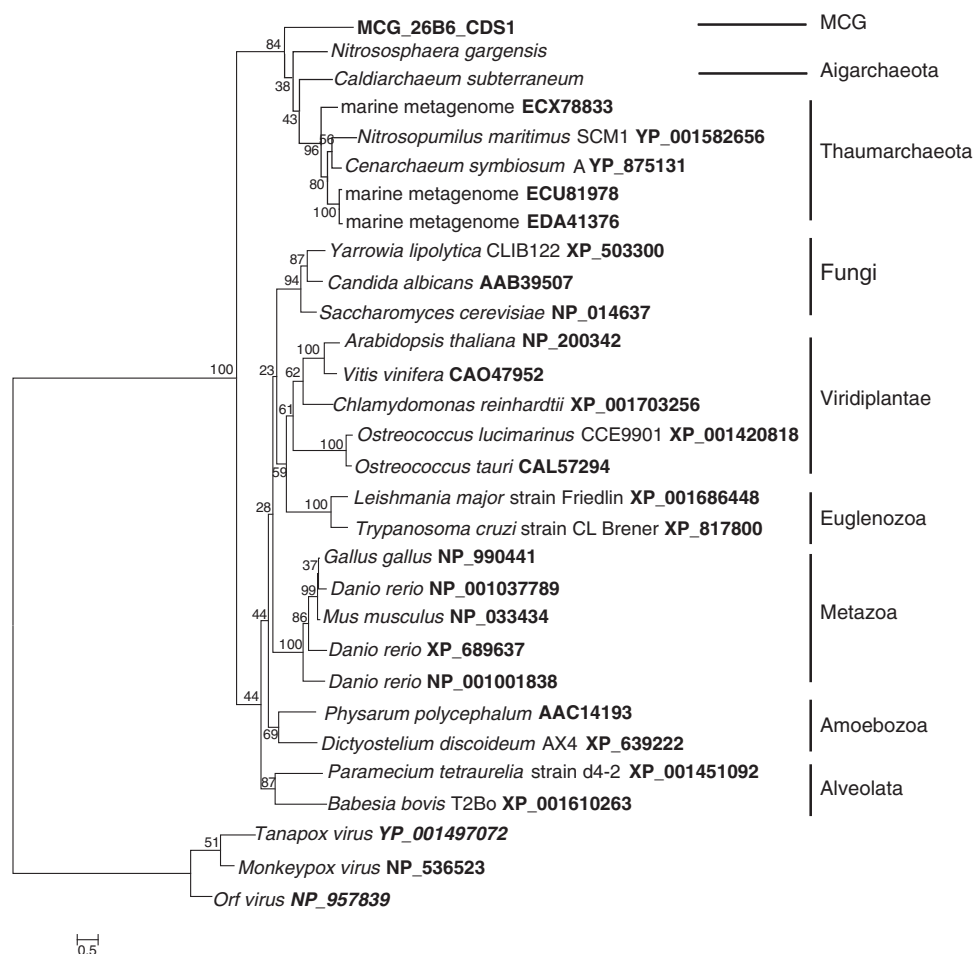
Figure 4 Unrooted maximum likelihood phylogenetic tree of TopoIB. TopoIB sequences from Thaumarchaeota, Virus and Euryotes. The numbers at the branches represent the bootstrap proportions. The scale bar represents the average number of substitutions per site.

of MCG requires isolates and more genomes of MCG members in the future.

Genes for aromatic compound degradation and expression verification

Within the genome fragment of 75G8, one CDS, CDS21, shared the highest identity (43% protein identity, *e*-value = 4e-90) with a methyl-accepting chemotaxis protein (MCPs) from *Nitrobacter winogradskyi* Nb-255 (Supplementary Table S1). MCPs are a family of receptors that mediate chemotaxis toward diverse signals, allowing organisms to respond to changes in the concentration of attractants and repellents in the environment by altering swimming behavior (Szurmant and Ordal, 2004). Two continuous CDSs (CDS16 and CDS17), locating upstream of MCPs protein in the MCG clone 75G8, were identified as putative 4-carboxymuconolactone decarboxylases (CMD) as matched to protein family HMM PF02627 (Supplementary Table S1). CMD catalyze the third step in the catabolism of protocatechuate (and therefore the fourth step in the catabolism of para-hydroxybenzoate, of 3-hydroxybenzoate, of vanillate and other compounds).

CMDs catalyze the decarboxylation of carboxymuconolactone, yielding β-ketoadipate enol-lactone, in the catabolism of aromatic compounds through the protocatechuate branch of the β-ketoadipate pathway (Stanier and Ornston, 1973).

On fosmid clone 75G8, putative CMD genes involved in protocatechuate catabolism locates at the upstream of the MCP gene. These putative MCG-CMD proteins show highest sequence identity to putative CMDs from bacterial strains (Supplementary Table S1 and Supplementary Figure S2). The CMD genes are widely distributed in various bacteria but rarely found in the available archaeal genomes (the presence of the CMD gene has only been observed in *Sulfolobus* and *Methanomicrobia* so far). The MCG-CMD seems to have a bacterial origin (Supplementary Figure S2). As there is still very limited information available, no confirmed conclusion could be obtained. Nevertheless, the presence of genes for both CMD and MCP in a genome fragment strongly suggests that MCG members (here, the MCG-G subcluster) may have the ability to utilize aromatic compounds. To test this hypothesis, we performed a substrate feeding cultivation experiment in which the source
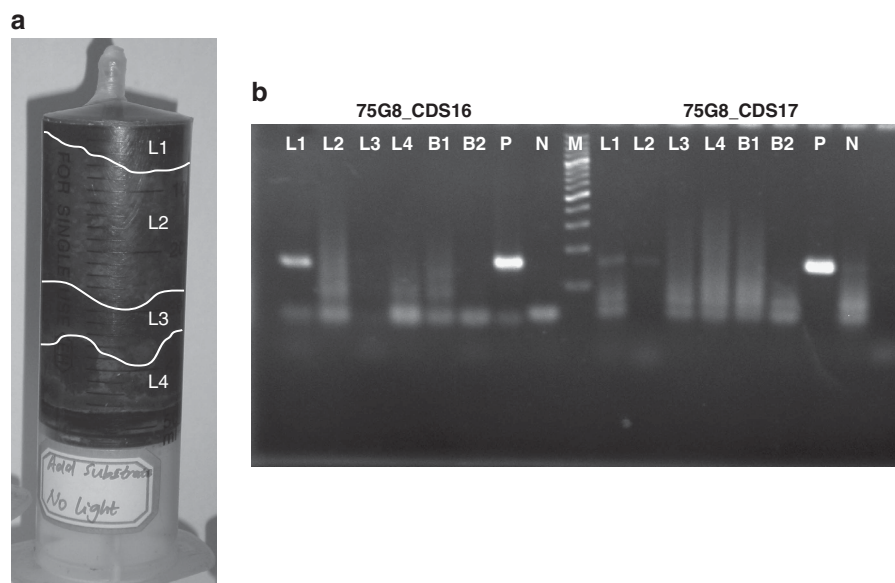
Figure 5 (**a**) A photo of the syringe filled with sediment and protocatechuate after 45 days' culturing. The labels from top to bottom (L1, L2, L3 and L4) correspond to the colors of the sediment layers. (**b**) Reverse transcription PCR (RT-PCR) analysis of 75G8_CDS16 and 75G8_CDS 17 from the RNA extracted from different layers of a culturing syringe (L1–L4), original sediment (B1) and a control sample (B2). L1–L4 indicate the different layers of the syringe sediment. B1 represents the original sediment sample without any treatment and B2 represents the control sample that was cultured under the same conditions but without protocatechuate. P indicates the positive control that used 75G8 fosmid DNA as the PCR template, and N indicates the negative control that used water as template.

sediment was supplemented with protocatechuate as substrate. Fresh sediment from the mangrove reservation district, the same location where the sediment collected for metagenome construction, was sampled by a syringe as shown in Figure 5a. Then, dissolved protocatechuate solution was injected into the syringe from the top hole and seeped into the syringe core (See Materials and Methods for details). The syringe was then sealed and incubated in a thermostatic room (26 °C) for 45 days. The incubated syringe was cut into four portions according to the color stratification (Figure 5a). Total RNA was extracted to examine the expression of MCG-CMD genes (CDS 16 and CDS 17). The expression of CDS16 was clearly observed only in the protocatechuate-supplemented sediment layer L1, whereas the expression of CDS17 was observed in both L1 and L2. In contrast, no expression of either CDS16 or CDS17 was observed in the original sediment sample or the control sample without protocatechuate supplementation (Figure 5b and Materials and Methods). The PCR bands of these two MCG-CMD genes were recovered from the gel, cloned and sequenced, and these sequences are identical to the CMD sequences in 75G8 genomic fragment. Therefore, the expression of CMD genes was stimulated by protocatechuate. The result of this preliminary substrate feeding experiment and gene expression analysis strongly supported our hypothesis that MCG archaea could utilize protocatechuate as a substrate. A previous study has considered MCG archaea to be heterotrophic anaerobes on the basis of depleted level of the stable ^{13}C (− 15 to − 28%) in whole archaeal

cells and intact archaeal membrane lipids (Biddle et al., 2006). Recently, one of the MCG group members was suggested to have the capability of protein degradation (Lloyd et al., 2013). Here, we identified another putative substrate: protocatechuate. It seemed to be very likely that MCG members within different MCG subgroups may have divergent substrate-utilizing capabilities, considering that MCG subgroups had extremely high genomic diversities as demonstrated above. Identifying other possible substrates in addition to protocatechuate by stable isotope probing and/or single-cell sequencing, represents a new exciting avenue of MCG research that may help elucidate the physiological properties of these organisms and facilitate isolation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

References

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Biddle J, Lipp J, Lever M, Lloyd K, Sorensen K, Anderson R *et al.* (2006). Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc Natl Acad Sci USA* **103**: 3846–3851.

Brochier-Armanet C, Gribaldo S, Forterre P. (2008). A DNA topoisomerase IB in Thaumarchaeota testifies for the presence of this enzyme in the last common ancestor of Archaea and Eucarya. *Biol Direct* **3**: 54.

DeLong E. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89**: 5685–5689.

Forterre P, Gribaldo S, Gadelle D, Serre M. (2007). Origin and evolution of DNA topoisomerases. *Biochimie* **89**: 427–446.

Fry J, Parkes R, Cragg B, Weightman A, Webster G. (2008). Prokaryotic biodiversity and activity in the deep subseafloor biosphere. *FEMS Microbiol Ecol* **66**: 181–196.

Inagaki F, Nunoura T, Nakagawa S, Teske A, Lever M, Lauer A *et al.* (2006). Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proc Natl Acad Sci USA* **103**: 2815–2820.

Inagaki F, Suzuki M, Takai K, Oida H, Sakamoto T, Aoki K *et al.* (2003). Microbial communities associated with geological horizons in coastal subseafloor sediments from the sea of okhotsk. *Appl Environ Microbiol* **69**: 7224–7235.

Jiang L, Zheng Y, Chen J, Xiao X, Wang F. (2011). Stratification of Archaeal communities in shallow sediments of the Pearl River Estuary, Southern China. *Antonie Van Leeuwenhoek* **99**: 739–751.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

Kubo K, Lloyd K, Biddle J, Amann R, Teske A, Knittel K. (2012). Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J* **6**: 1949–1965.

Le SQ, Gascuel O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol* **25**: 1307–1320.

Li PY, Xie BB, Zhang XY, Qin QL, Dang HY, Wang XM *et al.* (2012a). Genetic structure of three fosmid-fragments encoding 16S rRNA genes of the Miscellaneous Crenarchaeotic Group (MCG): implications for physiology and evolution of marine sedimentary archaea. *Environ Microbiol* **14**: 467–479.

Li Q, Wang F, Chen Z, Yin X, Xiao X. (2012b). Stratified active archaeal communites in the sediments of Jiulong River estuary, China. *Front Microbiol* **3**: 311.

Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD *et al.* (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496**: 215–218.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Lukashin AV, Borodovsky M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107–1115.

Meng J, Wang F, Zheng Y, Peng X, Zhou H, Xiao X. (2009). An uncultivated crenarchaeota contains functional bacteriochlorophyll a synthase. *ISME J* **3**: 106–116.

Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H *et al.* (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* **39**: 3204–3223.

Parkes RJ, Webster G, Cragg BA, Weightman AJ, Newberry CJ, Ferdelman TG *et al.* (2005). Deep sub-seafloor prokaryotes stimulated at interfaces over geological time. *Nature* **436**: 390–394.

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

Petersen TN, Brunak S, von Heijne G, Nielsen H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**: 785–786.

Sørensen KB, Teske A. (2006). Stratified communities of active Archaea in deep marine subsurface sediments. *Appl Environ Microbiol* **72**: 4596–4603.

Sonnhammer EL, Eddy SR, Durbin R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**: 405–420.

Stanier R, Ornston L. (1973). The beta-ketoadipate pathway. *Adv Microb Physiol* **9**: 89–151.

Szurmant H, Ordal W. (2004). Diversity in chemotaxis mechanisms among the bacteria and archaea. *Microbiol Mol Biol Rev* **68**: 301–319.

Teske A. (2006). Microbial communities of deep marine subsurface sediments: molecular and cultivation surveys. *Geomicrobiol J* **23**: 357–368.

Teske A, Sørensen K. (2008). Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME J* **2**: 3–18.

Zdobnov EM, Apweiler R. (2001). InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847–848.

Zhang X, Xu J, Xiao J, Xu J. (2009). Archaeal lineages detected in mangrove sediment in Zhangjiang estuary Fujian province. *J Shandong University (Natural Science)* **44**: 1–7. In Chinese.

Zhou J, Bruns M, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**: 316–322.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)