ORIGINAL ARTICLE

Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions

Ana-Belen Martin-Cuadrado¹, Francisco Rodriguez-Valera¹, David Moreira², José C Alba¹, Elena Ivars-Martínez¹, Matthew R Henn³, Emmanuel Talla⁴ and Purificación López-García²

¹*Division of Microbiology, Universidad Miguel Hernandez, Alicante, Spain;* ²*Unité d'Ecologie, Systématique et Evolution, CNRS UMR8079, Université Paris-Sud 11, Orsay, France;* ³*The Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA and* ⁴*Laboratoire de Chimie Bactérienne, Université de la Méditerranée Aix-Marseille II, CNRS-UPR9043, Marseille, France*

Marine planktonic archaea are widespread and abundant in deep oceanic waters but, despite their obvious ecological importance, little is known about them. Metagenomic analyses of large genome fragments allow access to both gene content and genome structure from single individuals of these cultivation-reluctant organisms. We present the comparative analysis of 22 archaeal genomic clones containing 16S rRNA genes that were selected from four metagenomic libraries constructed from meso- and bathypelagic plankton of different oceanic regions (South Atlantic, Antarctic Polar Front, Adriatic and Ionian Sea; depths from 500 to 3000 m). We sequenced clones of the divergent archaeal lineages Group 1A (Crenarchaeota) and Group III (Euryarchaeota) as well as clones from the more frequent Group I Crenarchaeota and Group II Euryarchaeota. Whenever possible, we analysed clones that had identical or nearly identical 16S rRNA genes and that were retrieved from distant geographical locations, that is, that defined pan-oceanic operational taxonomic units (OTUs). We detected genes involved in nitrogen fixation in Group 1A Crenarchaeota, and genes involved in carbon fixation pathways and oligopeptide importers in Group I Crenarchaeota, which could confirm the idea that these are mixotrophic. A two-component system resembling that found in ammonia-oxidizing bacteria was found in Group III Euryarchaeota, while genes for anaerobic respiratory chains were detected in Group II Euryarchaeota. Whereas gene sequence conservation was high, and recombination and gene shuffling extensive within and between OTUs in Group I Crenarchaeota, gene sequence conservation was low and global synteny maintained in Group II Euryarchaeota. This implies remarkable differences in genome dynamics in Group I Crenarchaeota and Group II Euryarchaeota with recombination and mutation being, respectively, the dominant genome-shaping forces. These observations, along with variations in GC content, led us to hypothesize that the two groups of organisms have fundamentally different lifestyles.

The ISME Journal (2008) 2, 865–886; doi:10.1038/ismej.2008.40; published online 8 May 2008

Subject Category: integrated genomics and post-genomic approaches in microbial ecology

Keywords: recombination; mutation; anaerobic respiration; carbon fixation; N metabolism; GC content

Introduction

Deep-sea microbiology is still poorly understood, despite the fact that the ocean is one of the largest biotopes on Earth and critically important for global nutrient cycling. Molecular diversity approaches based on the amplification, cloning and sequencing of 16S rRNA genes have revealed a variety of deep-sea archaeal, bacterial and picoeukaryotic lineages

Correspondence: P López-García, Unité d'Ecologie, Systématique et Evolution, CNRS UMR8079, Université Paris-Sud, bâtiment 360, Orsay Cedex 91405, France.

E-mail: puri.lopez@u-psud.fr

Received 29 January 2008; revised 18 March 2008; accepted 25 March 2008; published online 8 May 2008

that are often very divergent from cultivated species. Analyses of metagenomic libraries from marine, including deep-sea, plankton provide a powerful strategy for obtaining additional information about the structure, ecology and evolution of microbial communities inhabiting the marine ecosystem by allowing access to genes of as yet uncultured microorganisms (Stein *et al.*, 1996; Beja *et al.*, 2000a,b, 2002a,b; López-García *et al.*, 2004; Venter *et al.*, 2004; DeLong *et al.*, 2006; Martín-Cuadrado *et al.*, 2007; Rusch *et al.*, 2007).

Archaea have been detected throughout the oceanic water column and are quantitatively important members of picoplankton in the deep ocean, which suggests that organisms from this domain might have a fundamental role on global biogeochemical cycles (Karner *et al.*, 2001; Francis *et al.*, 2005; Herndl *et al.*, 2005). Pelagic microbial communities include psychrotolerant, free-living, as yet uncultivated marine archaea, and might also include anaerobic, particle-attached, methanogenic archaea. So far marine archaea identified in planktonic 16S rRNA gene libraries or metagenomic libraries fall into one of the following four lineages: Group I within the Crenarchaeota, and Groups II–IV within the Euryarchaeota (DeLong, 1992; Murray *et al.*, 1998; López-García *et al.*, 2001; Bano *et al.*, 2004). Recently, a new subgroup of crenarchaeota distantly related to the extensive Group I was identified, the so-called Group 1A, which clustered some fosmid sequences from the Pacific ALOHA water column (DeLong *et al.*, 2006) with the environmental clone pSL12 from a Yellowstone hot spring (Barns *et al.*, 1996).

Being elusive to isolation and growth in pure culture, little is known about marine archaeal physiology, and only two complete genomes from non-thermophilic crenarchaeota are available to date, that of *Nitrosopumilus maritimus* SCM1, and that of the sponge symbiont *Cenarchaeum symbiosum* A (Hallam *et al.*, 2006a). *N. maritimus*, which was isolated from a laboratory fish tank to become the first cultivated representative of Group I Crenarchaeota, grows by aerobically oxidizing ammonium to nitrite, a metabolic ability previously unknown for the archaea (Konneke *et al.*, 2005). The abundance of Group I archaeal cells in the deep ocean paralleled with significant expression levels of archaeal genes putatively encoding the ammonia monooxygenase subunit A (*amoA*) strongly supports the interpretation that at least a large fraction of the marine crenarchaeota are nitrifiers (Wuchter *et al.*, 2006; Lam *et al.*, 2007). Given the Earth's volume occupied by deep oceans, this would render archaea critical players in the N cycle at a global scale. Furthermore, deep-sea ammonium-oxidizing archaea may also be ecologically important for carbon cycling, as they appear to be autotrophic (Ingalls *et al.*, 2006) or at least mixotrophic. Indeed, *C. symbiosum* possesses the enzyme genes for the newly described 3-hydroxypropionate/4-hydroxy-

butyrate autotrophic CO₂ fixation pathway (Hallam *et al.*, 2006a,b; Berg *et al.*, 2007) and the Global Ocean Sampling database contains high proportions of genes for one key enzyme in this route, suggesting that this archaea-specific pathway is widely distributed in oceans (Berg *et al.*, 2007; Rusch *et al.*, 2007). However, although shot-gun sequencing approaches (for example, Global Ocean Sampling analyses) allow the association of functional genes to given environments, linking metabolic genes to specific phylogenetic lineages directly is only possible for very low-diversity biotopes (Tringe *et al.*, 2005). By contrast, sequencing of large genomic clones from metagenomic libraries allows assigning functional genes to particular lineages provided that appropriate phylogenetic markers colocalize in the same genome fragment. Through this approach proteorhodopsin-like genes, possibly acquired by horizontal gene transfer from phototrophic bacteria (Beja *et al.*, 2000a), were found in *bona fide* Group II Euryarchaeota inhabiting photic layers, though not in euryarchaeota from deeper waters (Frigaard *et al.*, 2006).

The structure and gene content of a limited number of sequenced large metagenomic clones carrying ribosomal RNA genes affiliated to marine Group I Crenarchaeota and Group II Euryarchaeota have been published to date (Stein *et al.*, 1996; Schleper *et al.*, 1998; Beja *et al.*, 2000b, 2002a; López-García *et al.*, 2004; Moreira *et al.*, 2004; Frigaard *et al.*, 2006). These studies were fundamentally descriptive; a few of them compared the sequences of two genomic fragments containing identical or nearly identical 16S rRNA sequences. The first of them compared two strains of *C. symbiosum*, which revealed an unexpected degree of genomic variation that might imply considerable functional diversity within single populations of coexisting microbial strains (Schleper *et al.*, 1998, 2005). Likewise, a pioneer comparison of two large genomic fragments of planktonic crenarchaeota from Antarctic surface water (fosmid 74A4) and North Pacific 200-m deep plankton (fosmid 4B7) reported significant genomic divergence between fragments with relatively similar 16S rRNA gene sequences (Beja *et al.*, 2002a). This suggested that this phylogenetic marker might not reflect the actual genomic and physiological diversity underlying the species or operational taxonomic unit (OTU) level, which is usually considered to be at ≥ 97 –98% nucleotide identity at this locus.

Comparing genomic sequences of the same OTUs from different geographic locations or from samples characterized by different environmental parameters can reveal interesting information about the microevolution and ecological adaptations of marine archaea. Yet genomic data for deep-sea Group I Crenarchaeota and Group II Euryarchaeota are rather poor, with most of the genomic information available (including the two complete genome sequences) corresponding to DNA samples or

isolates from the photic zone. Furthermore, so far there is no genomic information about members of Euryarchaeota Groups III and IV and Crenarchaeota Group 1A. In an effort to generate additional information about marine deep-sea planktonic archaea, we have sequenced and analysed 22 archaeal genome fragments (~770 Mb) identified in metagenomic fosmid libraries of deep plankton from geographically distant and very different oceanic water masses: Ionian Sea (3000 m), Adriatic Sea (1000 m), South Atlantic (1000 m) and Antarctic Polar Front (500 m). We present the first genomic data about Group 1A Crenarchaeota and Group III Euryarchaeota. Comparative genomic analyses of Groups I and II genomic fragments belonging to the same OTUs (based on 16S rDNA similarity) from different geographical locations suggest markedly different genome dynamics between members of these two archaeal groups.

Materials and methods

Sample collection and metagenomic library construction

The characteristics of the metagenomic planktonic libraries from which a selection of archaeal genomic clones are analysed in this study are summarized in Table 1. The 500 m-deep Polar Front cosmid library (DeepAnt) was constructed in two sequential steps yielding, initially, 6107 clones (López-García *et al.*, 2004) and later extended by 3200 additional clones (Moreira *et al.*, 2006). The 3000-m deep Ionian Sea fosmid library (KM3) was constructed as described previously (Martin-Cuadrado *et al.*, 2007). For the 1000-m deep South Atlantic library (SAT1000), 100 l sea water was collected in December 1998 during

the oceanographic cruise DHARMA98 and for the 1000-m deep Adriatic library (AD1000), 200 l was collected during the 2003 May Urania cruise in the Gulf of Manfredonia. In both cases, sea water was collected with a CTD rosette and sequentially filtered on board through a 5 µm pore size polycarbonate filter and through 0.22-µm pore size Sterivex filters (Durapore; Millipore, Billerica, MA, USA) using a peristaltic pumping system. Sterivex filters retaining the 0.2–5 µm diameter planktonic cells were conserved in lysis buffer (40 mM EDTA, 50 mM Tris/HCl, 0.75 M sucrose) at –20 °C until DNA extraction. Filters were thawed on ice and then treated with 1 mg ml^{–1} lysozyme and 0.2 mg ml^{–1} proteinase K (final concentrations). Nucleic acids were extracted with phenol–chloroform–isoamyl alcohol and chloroform–isoamyl alcohol, and then concentrated using a microconcentrator (Centricon 100; Amicon, Millipore, Billerica, MA, USA). DNA integrity was checked by agarose gel electrophoresis. Fosmid genomic libraries were constructed from approximately 0.2 µg (SAT1000) and 10 µg (AD1000) of DNA from the 0.2- to 5-µm plankton fraction using the CopyControl Fosmid Library Production Kit (Epicentre) as described by the manufacturer's instructions yielding 5413 and 38 704 clones, respectively. All fosmid libraries were constructed and stored at the University of Paris-Sud, Orsay.

Screening of libraries and sequencing of genomic clones

The DeepAnt and KM3 libraries were screened for the presence of archaeal 16S rRNA genes as previously described (López-García *et al.*, 2004; Martin-Cuadrado *et al.*, 2007). Clones from the AD1000 and SAT1000 libraries were pooled in

Table 1 Characteristics of sampling sites and planktonic metagenomic libraries used in this study

Name of library	KM3	AD1000	SAT1000	DeepAnt
Number of clones	20 757	38 704	5413	9307
Library size (Mbp) ^a	730	1360	190	330
Plankton fraction size (µm)	0.2–5	0.2–5	0.2–5	0.2–5
Amount of DNA used (µg)	~1	~10	~0.2	~0.2
Oceanic region	Ionian Sea	Adriatic Sea	South Atlantic	Antarctic Polar Front
Sampling site coordinates	36°20'N; 15°39'E	41°36'N; 17°22'E	56°18'S; 57°39'W	59°19'S; 55°45'W
Collection date	November 2004	May 2003	December 1998	December 1998
Collection depth (m)	3000	1000	1000	500
Sea bottom depth (m)	3243	1099	4320	3700
Temperature (°C)	13.4	12.5	0.3	0.4
Salinity (‰)	38.7	38.6	34.4	34.5
Dissolved oxygen (ml l ^{–1})	4.5	5.2	5.0	4.8
Reference	Martin-Cuadrado <i>et al.</i> (2007)	This study	This study	López-García <i>et al.</i> (2004); Moreira <i>et al.</i> (2006)

^aConsidering an average insert size of 35 kb.

groups of 96, and DNA was extracted from pooled cultures using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA, USA). The libraries were PCR-screened for the presence of archaeal 16S rRNA genes using combinations of the archaea-specific forward primers 21F (5'-TTCCGGTTGATCCTGCCGGA), Ar109 (5'-AC(G/T)GCTGCTCAGTAACACGT), ANMEF (5'-GGCTCAGTAACACGTGGA) and the prokaryotic-specific reverse primer 1492R (5'-GGT TACCTTGTTACGACTT). PCR reactions were carried out under the following standard conditions: 35 cycles (denaturation at 94 °C for 15 s, annealing at 50 °C for 30 s, extension at 72 °C for 2 min) preceded by 2 min denaturation at 94 °C and followed by 7 min extension at 72 °C. PCR products were partially sequenced. A total of 28, 30 and 9 archaeal 16S rDNA sequences were identified in the KM3, AD1000 and Sat1000 libraries, respectively. After partial sequencing and preliminary phylogenetic analyses, 12 fosmid clones affiliating to crenarchaeota and 9 to euryarchaeota were selected for complete sequencing. Sequencing was carried out at the Broad Institute at the Massachusetts Institute of Technology thanks to the Betty and Gordon Moore Foundation 'Marine Microbiology Initiative' (<http://www.moore.org/marine-micro.aspx>). An additional euryarchaeotal clone, DeepAnt-15E7, which we had identified and sequenced (Genome Express, Meylan, France) from a cosmid library of 500-m deep picoplankton at the Antarctic Polar Front (López-García *et al.*, 2004) was included in this analysis, making a total of 22 archaeal clones analysed (Table 2). All sequences were obtained in a single contig except KM3-136-D10 and SAT1000-49-D2, which were assembled in two different ones.

Annotation and analysis of genome fragments

Protein coding genes were predicted using the annotation package GLIMMER (Delcher *et al.*, 1999), SEEDs (Overbeek *et al.*, 2005) and were further manually curated. Spacers were subsequently searched against the non-redundant database (<http://www.ncbi.nlm.nih.gov/>) using BLAST (Altschul *et al.*, 1997) to ensure that no open reading frame (ORF) had been missed. Identified ORFs were compared to known proteins in the non-redundant database using BLASTX. All hits with an *E*-value greater than 10^{-5} were considered nonsignificant. COGNITOR was used for COG (Clusters of Orthologous Groups) assignments as well as COG functional categories (Tatusov *et al.*, 2001). tRNAs were identified using tRNAscan-SE (Lowe and Eddy, 1997). Putative protein transmembrane domains were predicted using TMHMM 2.0 (Krogh *et al.*, 2001). The GEECEE and CUSP programs from the EMBOSS package (Rice *et al.*, 2000) were used to calculate GC content and codon usage, respectively. For comparative analyses, reciprocal BLASTN and TBLASTXs searches between the different fosmids were carried out, leading to the identification of

regions of similarity, insertions and rearrangements. To allow the interactive visualization of genomic fragment comparisons, we used Artemis Comparison Tool ACTv.6 (Carver *et al.*, 2005) and MAUVE (Darling *et al.*, 2004). To identify possible recombination regions between fosmids, we applied different programs available in the RDP version program package (Martin *et al.*, 2005), including GENECONV (Sawyer, 1999), RDPv.beta14 and MaxChi (Posada and Crandall, 2001). The SimPlot graphic (Ray, 1998) shown in Figure 6 was done with the following parameters: window 1000 pb, step 20 pb, GapStrip: On, Kimura (two parameter) and T/t 2.0.

Phylogenetic analysis

Archaeal 16S rRNA gene sequences detected in genomic clones were aligned using MUSCLE (Edgar, 2004) with their closest relatives in databases as identified by BLAST (Altschul *et al.*, 1997), those from the ALOHA water column (DeLong *et al.*, 2006) and those from available Group I crenarchaeotal genomes and selected Groups I and II archaeal genome fragments. Alignments were manually edited using the ED program of the MUST package (Philippe, 1993). Gaps and ambiguously aligned positions were excluded from our analyses, yielding 1184 and 1118 positions available for constructing the trees of crenarchaeota and euryarchaeota, respectively. Maximum likelihood trees were reconstructed using TREEFINDER (Jobb *et al.*, 2004) applying a general time reversible model of sequence evolution, and taking among-site rate variation into account by using a four-category discrete approximation of a Γ distribution plus invariant sites. ML bootstrap proportions were inferred using 1000 replicates. Phylogenetic trees were viewed using the program TREEVIEW (Page, 1996). For the intergenic transcribed spacer (ITS) analysis (Figure 2b), a neighbour-joining tree was constructed using nearly all positions (159), including gaps to maximize the signal, using the programs available in the MUST package. Protein sequence alignments were generated with MUSCLE and ClustalW (Chenna *et al.*, 2003) and edited manually as necessary. Gblocks was used to eliminate ambiguously aligned positions (Castresana, 2000). Phylogenetic analyses were then performed using the MEGA4 phylogenetic tool software package (Zmasek and Eddy, 2001; Kumar *et al.*, 2004). To construct the concatenate of the euryarchaeotal Group II ribosomal proteins only L3, L4, L23, L2/LB, S19, L22, S3, L29, S5, L30 and L15 were used to include KM3-136-D10 in the analysis.

Accession numbers

The sequences have been deposited in GenBank under the accession numbers AC215088–AC215101, AC215103–AC215109 and EU556724.

Table 2 General information about the archaeal genome fragments analysed in this study from various metagenomic libraries

Name of genomic clone	% GC	Assembly coverage	Insert length (bp)	No. predicted ORFS	No. hypothetical proteins	Average ORF length (bp)	Coding regions (%)	Closest matches to 16S rRNA genes in genomic clones			Closest cultivated archaeon (% identity at 16S rDNA locus)
								Clone (accession no.)	Isolation source	% id.	
Crenarchaeota group I											
AD1000-56-E4	35.23	20.09	35 441	33	13	763	83.9	Sd-EA04 (AB194000)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Cenarchaeum symbiosum strain B (92)
AD1000-202-A2	35.59	21.28	38 090	37	20	776	87.1	MaSc-NEA03 (AB193971)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	98	Candidatus Nitrosopumilus maritimus (92)
AD1000-207-H3	34.48	18.98	41 899	42	17	739	85.1	MaSc-NEA01 (AB193956)	Mariana Trough hydrothermal vent water (0.2 µm filtrate)	98	Cenarchaeum symbiosum strain B (92)
KM3-34-D9	37.88	9.65	22 647	18	6	880	89.7	MaSc-NEA03 (AB193971)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Candidatus Nitrosopumilus maritimus (92)
KM3-47-D6	37.11	17.98	39 182	33	9	950	93.9	MaSc-NEA01 (AB193956)	Mariana Trough hydrothermal vent water (0.2 µm filtrate)	99	Cenarchaeum symbiosum strain B (92)
KM3-86-C1	36.31	15.16	30 101	29	10	711	83.7	SSM263-NA01 (AB193964)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Cenarchaeum symbiosum strain B (92)
SAT1000-21-C11	35.49	22.91	36 490	44	20	645	88.1	SSM263-NA01 (AB193964)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Cenarchaeum symbiosum strain B (92)
SAT1000-23-F7	36.04	13.59	37 834	36	19	763	84.4	MaSc-NEA03 (AB193971)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Candidatus Nitrosopumilus maritimus (92)
SAT1000-49-D2 ^a	35.27	8.04	35 428	38	17	670	86.4	MaSc-NEA01 (AB193969)	Mariana Trough hydrothermal vent water (0.2 µm filtrate)	99	Cenarchaeum symbiosum strain B (92)
Crenarchaeota group 1A											
AD1000-23-H12	34.31	12.13	27 960	27	7	826	88.6	ALOHA200m_0002 (EF106844)	North Pacific ALOHA Station 200 m depth	99	Candidatus Nitrosopumilus maritimus (82)
AD1000-325-A12	33.53	20.13	36 134	31	6	931	91.5	Sd-NA (AB193977)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	95	Candidatus Nitrosopumilus maritimus (82)
KM3-153-F8	33.76	20.26	37 751	32	5	967	94.0	Sd-NA (AB193977)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	93	Candidatus Nitrosopumilus maritimus (82)
Euryarchaeota group II											
AD1000-18-D2	58.85	26.04	37 814	29	5	1147	91.9	SSM263-NA03 (AB193966)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	98	Methanococcus maripaludis C7 (77)
DeepAnt-15E7	56.04	5.1	38 890	40	2	856	92.7	Sd-EA02 (AB193998)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	Aciduliprofundum boonei T469 (79)
KM3-72-G3	58.71	24.13	30 340	26	7	1042	94.2	Sd-EA02 (AB193998)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	98	Picrophilus torridus DSM 9790 (78)

Table 2 Continued

Name of genomic clone	% GC	Assembly coverage	Insert length (bp)	No. predicted ORFs	No. hypothetical proteins	Average ORF length (bp)	Coding regions (%)	Closest matches to 16S rRNA genes in genomic clones			Closest cultivated archaeon (% identity at 16S rDNA locus)
								Clone (accession no.)	Isolation source	% id.	
KM3-85-F5	59.97	26.7	38 426	40	5	846	91.9	Scd-NEA03 (DQ299286)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	96	<i>Methanothermobacter thermautotrophicus</i> ΔH (78)
KM3-130-D10	59.61	23.33	40 302	43	5	817	87.2	Scd-NEA03 (AB193995)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	<i>Methanococcus maripaludis</i> S2 (77)
KM3-136-D10 ^a	45.31	8.62	17 287+8121	15+5	0+1	1082	89.1	CTD005-13A (AY856358)	Deep-sea water near Baby Bare Seamount 2540 m	98	<i>Methanococcoides burtonii</i> DSM 6242 (77)
SAT1000-15-B12	58.86	15.98	35 919	36	3	899	94.2	Scd-NEA03 (AB193966)	Suiyo Seamount hydrothermal vent water (0.2 µm filtrate)	99	<i>Methanothermobacter thermautotrophicus</i> ΔH (78)
<i>Euryarchaeota</i> group III AD1000-40-D7	38.64	24.55	35 669	23	5	1132	89.6	CTD005-33A (AY856361)	Deep-sea water near Baby Bare Seamount, 2540 m depth, Northeast Pacific Ocean	98	<i>Methanococcus vannielii</i> SB (78)
KM3-28-E8	54.83	25.04	38 847	28	1	1167	87.9	CTD005-16A (AY856362)	Deep-sea water near Baby Bare Seamount, 2540 m depth, Northeast Pacific Ocean	97	<i>Methanococcus vannielii</i> SB (78)
SAT1000-53-B3	40.66	9.44	31 425	23	3	1216	93.4	CTD005-33A (AY856361)	Deep-sea water near Baby Bare Seamount, 2540 m depth, Northeast Pacific Ocean	98	<i>Methanococcoides burtonii</i> DSM 6242 (79)

^aFosmids assembled initially in two contigs; SAT1000-49-D2 was assembled in a single contig by a stretch of 650 'N'.
id., identity.

Results

Archaeal fosmids in deep-sea metagenomic libraries

To gain insights into the genomic diversity of deep-sea marine archaea, we screened four metagenomic libraries (Table 1) of meso- and bathypelagic plankton (size fraction 0.2–5 µm) from various oceanic locations looking for the presence of genomic clones containing archaeal 16S rRNA genes, a selection of which were completely sequenced and analysed. Two of the metagenomic libraries had been constructed previously: the cosmid library DeepAnt (Antarctic Polar Front, 500 m depth) (López-García *et al.*, 2004; Moreira *et al.*, 2006) and the fosmid library KM3 (Ionian Sea, 3000 m depth) (Martín-Cuadrado *et al.*, 2007). For this study, we constructed and screened two additional fosmid libraries: SAT1000 (South Atlantic, 1000 m depth) and AD1000 (Adriatic Sea, 1000 m depth). With 9307, 20757, 5413 and 38704 genomic clones, respectively, that is, a total of ~74 000 genomic clones with an average insert size of 35 kb; this represents approximately 2.6 Gb of stored genetic information (Table 1).

These libraries were scanned by PCR using various 16S rDNA archaeal-specific primer combinations revealing a relatively high proportion of archaeal clones (López-García *et al.*, 2004; Martín-Cuadrado *et al.*, 2007 and this study). We obtained approximately 75 archaeal 16S rDNA sequences from all the libraries, which were compared by BLAST to sequences in databases and classified in different archaeal groups after a preliminary phylogenetic analysis (data not shown). Using the 16S rDNA-containing clones as a proxy, the relative abundance of different archaeal groups appeared to vary depending on the libraries (Figure 1). Although Group I Crenarchaeota are thought to constitute the majority of the archaeal fraction in deep waters (Karner *et al.*, 2001; DeLong *et al.*, 2006), our study revealed important differences among libraries, with Groups II and III Euryarchaeota dominating some of them. This might partly reflect biases in euryarchaeota detection due to differential primer/probe

specificity, a problem that is limited in the present study by the use of a variety of primer combinations. In any case, it certainly reflects real variation in archaeal community composition of different sea water masses and geographic locations. Since the Mediterranean basin is rather warm and with very particular physico-chemical characteristics (Table 1), different archaeal composition would be expected for the Adriatic and Ionian libraries compared to open ocean libraries. However, while crenarchaeota appeared dominant, though not overwhelmingly, in the Ionian KM3 library, they accounted for less than one-third of the archaea in the Adriatic one (Figure 1). Similarly the South Atlantic and Antarctic Polar Front libraries, from geographically close sampling stations (Table 1) albeit from quite different water masses (Supplementary Figure S1), exhibited an opposite dominance of archaeal groups, with crenarchaeota clearly dominating the SAT1000 library and Group II Euryarchaeota the DeepAnt one. Notably, the usually scarce Group III Euryarchaeota appeared to be the most represented euryarchaeotal clade in the South Atlantic library.

From the preliminary 16S rDNA archaeal phylogenetic tree, we selected 22 archaeal genomic clones for complete sequencing and comparative analysis based on two major criteria: (i) they represented lineages about which there was no genomic information available as yet (for example, Group 1A Crenarchaeota, Group III Euryarchaeota) or (ii) they had nearly identical 16S rDNA sequences and could therefore be ascribed to a single OTU (here defined as lines having 16S rDNAs with ≥98% identity) but came from distant oceanic locations (Table 2; Figures 2 and 3). Of the 22 selected fosmids, 12 belonged to the Crenarchaeota and, of these, 3 were more closely related to the recently identified Group 1A, initially defined by DeLong *et al.* (2006) on the basis of the environmental sequence pSL12 and the Pacific ALOHA fosmids HF770-041I11 and HF4000-039O16. However, the monophyly of the Group 1A with the addition of these environmental sequences remains only weakly supported, the marine planktonic sequences forming a clade distinct from the Yellowstone hot spring clones pSL12, OPPD032 and YNP-ObP-A62 (Figure 2a). The remaining nine crenarchaeotal fosmids belonged to the widespread Group I and comprised three distinct cosmopolitan OTUs (A, B and C) with, for each, one fosmid from the Adriatic, one from the Ionian and a third one from the South Atlantic (Figure 2a). In support of their pan-oceanic distribution, these three OTUs included environmental sequences from yet another oceanic basin, the Pacific Ocean, at the ALOHA water column (HF clones) and the Suiyo Seamount (SSM264-EA01, MaSc-NEA01 and MaSc-NEA03) (Figure 2a). From the 10 euryarchaeotal genomic clones sequenced, 3 were ascribed to the Group III and the rest to deep-sea Group II members, as none of the 16S rDNA sequences from our libraries branched with the phototrophic euryarchaeota

Figure 1 Relative proportion of archaeal genomic clones containing 16S rRNA genes in the metagenomic libraries from meso- and bathypelagic plankton used in this study.

Figure 2 (a) Maximum likelihood phylogenetic tree based on the rRNA genes contained in the fosmids of crenarchaeota analysed in this study. Fosmid clone names indicate the metagenomic library of provenance: KM3, Ionian Sea, 3000 m depth; AD1000, Adriatic Sea, 1000 m depth; SAT1000, South Atlantic, 1000 m depth; DeepAnt, Antarctic Polar Front, 500 m depth; HF, North Pacific followed by a number indicating their depth (DeLong *et al.*, 2006). Previously sequenced environmental fosmid clones are indicated as such, the rest of clone names correspond to environmental sequences, [accession number]. (b) Neighbour-joining tree of the intergenic transcribed spacer (ITS) sequences adjacent to the 16S rRNA gene in various crenarchaeotal fosmids.

(Frigaard *et al.*, 2006). The recently isolated *Aciduliprofundum boonei*, an obligate thermoacidophilic sulphur- or iron-reducing heterotrophic deep-sea vent archaeon representative of the DHVE2 group (Reysenbach *et al.*, 2006), branched closer to the

Thermoplasmatales than to the Groups II and III (Figure 3 and data not shown). As in the case of crenarchaeota, we sequenced fosmids from the same pan-oceanic OTUs coming from at least one Mediterranean and one South Atlantic or Antarctic location. Again, that these OTUs are cosmopolitan was further supported by the formation of tight clades with Pacific sequences from the ALOHA water column (HF clones), Suiyo Seamount hydrothermal sea water (SSM263-NA03) or ridge flank crustal fluids in the North East Pacific (pIVWA104, CTD005-33A) (Huber *et al.*, 2006) (Figure 3).

The archaeal genomic fragments sequenced ranged from 22.6 to 41.8 kb, with an average size of 35 kb (Table 2). All fosmids were assembled in a single contig with the exception of KM3-136-D10 and SAT1000-49-D2, which were assembled in two contigs. Through alignments between SAT1000-49-D2 and the other fosmids, we concluded that the most probable orientation of the two contigs was that shown in Figure 4, and we joined them artificially by a stretch of 650 'N' that most likely represented the region missed during sequencing. KM3-136-D10 appeared to lack a larger region, so that we kept the two contigs separated in our analyses. On average, gene density was relatively high, with >90–91% coding regions in all fosmids except those from Group I Crenarchaeota, which had slightly less dense coding regions (86.9% on average). These genome fragments were annotated and the gene content and organization compared within the different archaeal groups as follows.

Genomic organization and detection of high intra- and inter-genomic recombination levels in Group I Crenarchaeota

All Group I Crenarchaeota genome fragments had a very low GC content (Table 2), with an average of 35.9%, a value closer to that of the free-living *N. maritimus* SCM1 genome (33.9%) than to that of the symbiont *C. symbiosum* (57.4%). Genome organization was remarkably variable (Figure 4), despite the fact that the nine Group I crenarchaeotal fosmids sequenced formed three well-defined, though rather close (especially OTUs B and C), OTUs in terms of 16S rRNA phylogeny (Figure 2a). This variability was particularly evident at the level of gene and genome fragment rearrangements within and among OTUs (Figure 4), in spite of similar overall gene content, gene similarity (Supplementary Table S1) and small portions of co-linear regions being conserved. This is suggestive of a high level of intra-genomic recombination leading to gene shuffling.

Remarkably, ITS between the 16S and 23S rRNA genes, which are generally very variable, were extremely conserved. Thus, although they were rather short (much shorter than, and very different from, those of soil crenarchaeota), having 130 bp on average, their phylogenetic analysis allowed retrieving

Figure 3 Maximum likelihood phylogenetic tree based on the rRNA genes contained in the fosmids of euryarchaeota analysed in this study. Fosmid clone names indicate the metagenomic library of provenance: KM3, Ionian Sea, 3000 m depth; AD1000, Adriatic Sea, 1000 m depth; SAT1000, South Atlantic, 1000 m depth; DeepAnt, Antarctic Polar Front, 500 m depth; HF, North Pacific followed by a number indicating their depth (DeLong et al., 2006). Previously sequenced environmental fosmid clones are indicated as such, the rest of clone names correspond to environmental sequences, [accession number].

trees that were overall congruent with 16S rRNA trees even using simple *p*-distance-based trees considering all positions (including gaps) (Figure 2b and Supplementary Figure S1). This degree of conservation had been observed previously in environmental sequences (García-Martínez and Rodríguez-Valera, 2000) and the Group I crenarchaeote clone DeepAnt-EC39 (López-García et al., 2004). The most obvious difference with 16S rRNA trees was the split of OTU C, since the ITS sequence of the OTU C clone SAT1000-49-D2 was identical to, and branched with, those of OTU B (Figure 2 and Supplementary Figure S2). This behaviour was suggestive of a recombination event between members of these two OTUs. To detect other possible recombination events between the different OTU clones at the 16S-23S rDNA transition region, we made *p*-distance-based phylogenetic trees using different contiguous windows along an

alignment of the 16S rDNA + ITS + 23S rDNA including all positions (Supplementary Figure S2A). The tree obtained with the last 300 positions of the 16S rRNA gene was congruent with the maximum likelihood phylogenetic tree obtained for this whole gene (Figure 2a). OTUs A, B and C were clearly distinct and the clone DeepAnt-EC39 branched with moderate support at the base of OTU C. The tree obtained using the adjacent 189 positions corresponding to the ITS region showed, as already seen, the clustering of OTU C clone SAT1000-49-D2 with OTU B (100% identity). DeepAnt-EC39 displayed a longer branch within the OTU B, which might be explained by recombination with environmental lines (compare Figure 2b and Supplementary Figure S2A). The tree obtained using the contiguous first 300 positions of the 23S rRNA gene was surprising as it showed an extreme high conservation (100% identity) of this region within the three OTUs, A, B and C, DeepAnt-EC39 being now identical to members of the OTU C. This level of conservation is possibly not related to a particular essential function of this region of the 23S rRNA in Group I Crenarchaeota, as the phylogenetic distance between *N. maritimus* and the fosmid 74A4, which can be used as internal control in the analyses, remains approximately the same in all the trees made. Finally, the tree obtained with the following 300 positions in the 23S rRNA (positions 300–600) retrieved the same phylogenetic relationships as did the 16S rRNA partial and global trees and the rest of the 23S rRNA gene.

Taken together, these observations strongly suggest that this region of the *rrn* operon is a hotspot for inter-genomic homologous recombination. This would explain why the ITS, in principle not subjected to strong purifying selection, is so astonishingly conserved within and between different OTUs. Intra-chromosomal recombination at the *rrn* operon is unlikely since these organisms possibly bear a single *rrn* operon. Two elements support this idea. First, *C. symbiosum* and *N. maritimus* genomes harbour one single *rrn* operon and, second, though synteny was not perfectly conserved, there were more or less long syntenic regions involving all the Group I Crenarchaeota fosmids examined here, which would be extremely improbable if these crenarchaeota contained more than one *rrn* operon (Figure 4). Therefore, recombination likely takes place with exogenous genomic DNA from more or less closely related archaea.

Regarding protein gene content and organization, some protein genes were very well conserved in our fosmids. These included, immediately upstream the 16S rDNA, the genes encoding glutamate semialdehyde aminotransferase, a tetra-tricopeptide-repeat-rich protein (TPR) and an unknown conserved transmembrane protein (Figure 4 and Supplementary Table S1). Though not equally conserved, genes for a NADP(H)-dependent nitroreductase or the transcriptional regulator AsnC, followed by a thioredoxin

Figure 4 Comparative genomic organization of fosmids from crenarchaeota. A neighbour-joining tree relating the different 16S rRNA markers found in the respective fosmids is shown on the left as reference. Conserved genomic regions between fosmids are indicated by grey shaded areas, grey intensity being a function of sequence similarity by BLASTN. ORFs are numbered from left (ORF 1 in each case) to right as shown in Supplementary Table S1. Particular ORFs mentioned in the text are highlighted by a graphic code. Note that SAT1000-49-D2 is a composite of two contigs joined by a stretch of 650 nucleotides (labelled as 'N').

reductase were often found immediately downstream of the 23S rRNA gene. BLASTP comparisons of the proteins found in these Group I crenarchaeotal fosmids with the non-redundant database (*E*-value cutoff of $1e-5$) showed that most of them had best hits with proteins of *N. maritimus* SCM1 (56.8% of blasted proteins), *C. symbiosum* A (11.3%) and the previously described marine crenarchaeotal clones DeepAnt-EC39 (21.0%), 4B7 (5.7%), 74A4 (0.7%) as well as other uncultured crenarchaeota. Hits with thermophilic crenarchaeota (*Sulfolobus acidocaldarius* DSM 639) were observed only occasionally (0.6%). In 1.4% of the cases, the best hits were bacterial proteins (0.7% of them were indeed cyanobacteria) followed by a slightly less similar protein in the *N. maritimus* SCM1 or *C. symbiosum* A genomes. These observations might be consistent with previous analysis of DeepAnt-EC39 showing that some genes had probably been acquired by horizontal gene transfer, likely by an ancestor of Group I Crenarchaeota, from euryarchaeota or bacteria (including well-supported

cases of transfers from cyanobacteria) (López-García *et al.*, 2004).

Although all the genes found in our bathypelagic fosmids had orthologues in both *N. maritimus* SCM1 and *C. symbiosum* A, overall synteny was not maintained, except for the three genes upstream the 16S rDNA mentioned above. However, colinearity was preserved in clusters of genes that were positioned in varying locations along the fosmids suggesting the occurrence of frequent rearrangements and recombination (Figure 5). To explore further the occurrence of recombination events at protein coding genes in our Group I crenarchaeotal fosmids, we looked for potential recombination spots in conserved syntenic regions within the different OTUs (Supplementary Tables S2 and S3). In the case of OTU A comprising fosmids AD1000-202-A2, KM3-34-D9 and SAT1000-23-F7, we aligned an overall syntenic region of 19 652 bp corresponding to ORFs 22–36 in SAT1000-23-F7, which included the glutamate semialdehyde aminotransferase, TPR and adjacent hypothetical

Figure 5 Regions of co-linearity shared by some of the deep-sea crenarchaeota fosmids studied here and the genome of *Nitrosopumilus maritimus* and crenarchaeotal genomic clones 4B7 and 74A4.

protein conserved genes as well as the translation elongation factor 2 (EF-2), a glucose dehydrogenase, a nitroreductase, two transcriptional regulators and seven additional conserved hypothetical protein genes (Supplementary Table S1). The Mediterranean KM3-34-D9 and the South Atlantic SAT1000-23-F7 had 93.85% nucleotide identity on average, whereas that between the two Mediterranean clones (Adriatic and Ionian) KM3-34-D9 and AD1000-202-A2 was slightly lower, 88.11%. These values supported the inclusion of these fosmids within a single OTU, in accordance with values expected for a single prokaryotic species (Konstantinidis and Tiedje, 2005). These data argue, again, in favour of true oceanic ubiquity for this OTU. We scanned this alignment with recombination breakpoint-detection methods and found up to eight potential sites (Supplementary Table S2). Among them, one of the most evident inter-genomic recombination events leading to genetic exchange within populations involved the EF-2 gene, as it exhibited a much higher similarity with its homologue in AD1000-202-A2 (96.2% nucleotide identity) than to KM3-34-D9 (75.3%) in contrast to the rest of the aligned region (Figure 6).

Considering together the six genomic clones that defined the closely related OTUs B and C, they

exhibited lower resemblance of gene organization among them than OTU A fosmids. This might imply a higher intra-genomic recombination rate, at least in the area surrounding the *rrn* operon, resulting in extensive genome reshuffling. For instance, in fosmid KM3-47-D6, synteny was restricted nearly exclusively to the genes upstream the *rrn* operon. Furthermore, 96% of the ORFs downstream of the *rrn* operon were not at all syntenic between OTU C clones KM3-47-D6 and AD1000-207-H3, and OTU B clones AD1000-56-E4 and SAT1000-21-C11. Nevertheless, we found, once again, clusters of genes that kept local synteny while being scrambled within and between members of OTUs B and C (Figure 4). We were able to align a fragment of 9663 bp corresponding to a syntenic region covering the ORFs 26–29 in SAT1000-49-D2, including also the genomic clone DeepAnt-EC39. Scanning this alignment with recombination breakpoint-detection programs, up to 9 potential recombination sites were identified (Supplementary Table S3). Inter-genomic recombination events might even involve members of more distant OTUs. Thus, one of the potential recombination events detected involved the NADPH nitroreductase of the OTU C fosmids KM3-47-D6 and SAT1000-49-D2, which showed an atypical clustering, being much more similar to that of the

Figure 6 Detection of recombination at the translation elongation factor 2 (EF-2) locus within the crenarchaeotal OTU A (SAT1000-23-F7, AD1000-202-A2 and KM3-34-D9). The curves show the similarity at nucleotide level of AD1000-202-A2 and KM3-34-D9 to the region comprised between open reading frames (ORFs) 22 and 36 in fosmid SAT1000-23-F7. The insets on the right show the phylogenetic relationships (neighbour-joining trees) retrieved within the OTU A from EF-2 compared to the concatenate of the different ORFs in this region.

OTU A fosmid KM3-34-D9 (96.2 and 95.8% nucleotide identity, respectively) than to any of their homologous counterparts in their respective OTUs. In addition, intra-genomic recombination can lead to the translocation of relatively big genome fragments. For example, a set of genes keeping strict or nearly strict synteny is observed downstream of the *rrn* operon in several, but not all, members of OTUs B and C but upstream the *rrn* operon in the sequences belonging to OTU A (boxed regions in Figure 4). This translocation is about 16 000 bp long and includes genes for two oligopeptide permeases, a helicase, the chaperone DnaJ, a membrane protein probably implicated in the uptake of heavy metals (DedA), two TPR repeat-containing proteins, a SAM domain protein, several hypothetical proteins and EF-2.

Taken together, all these observations provide strong support for the idea that these Group I archaeal genomes have high intra- and inter-genomic recombination rates, at least in the vicinity of the *rrn* operon. This conclusion was further strengthened by the observed ratio of protein genes having the same transcriptional sense as the *rrn* operon versus those with an opposite orientation. As can be seen in Figure 7, only 50–60% of protein genes in our Group I Crenarchaeota fosmids had the same orientation as the *rrn* operon. These values were strikingly different from those observed for the Group II Euryarchaeota genomic clones (see below) revealing different genome dynamics for these two archaeal groups.

Metabolic potential of bathypelagic Group I Crenarchaeota

Most of the genes identified encoded proteins involved in informational processes and many corresponded to conserved hypothetical proteins also present in other marine crenarchaeota (Supplementary Table S1). However, we also identified some genes that might provide clues as to the metabolic potential of these microbes. Thus, we found several genes encoding 3-hydroxypropionate cycle proteins involved in the putative carbon fixation pathway in the chemolithotrophic metabolism of marine crenarchaeota (Hallam *et al.*, 2006a,b). KM3-47-D6 contained a gene for a methylmalonyl-CoA mutase (*mcmA1*) and a methylmalonyl-CoA epimerase (*mce*), and KM3-86-C1 and SAT1000-49-D2 harboured an enoyl-CoA hydratase gene (*fadB*). In AD1000-202-A2 and KM3-34-D9, a region downstream of the *rrn* operon contained a gene for biotin ligase (*birA*), a protein required for assembly and activation of the carboxylase complex (Rodionov *et al.*, 2002). Next to *birA* genes was a succinyl-CoA synthetase (alpha subunit), an enzyme that mediates the formation of succinyl-CoA from succinate in the reductive tricarboxylic acid pathway. Though these genes are not the key enzymes of the reductive tricarboxylic acid pathway cycle, the 3-hydroxypropionate pathway, or its variant, the 3-hydroxypropionate/4-hydroxybutyrate autotrophic CO₂ assimilation pathway recently discovered in archaea, including *C. symbiosum* and *N. maritimus* (Berg *et al.*, 2007), their presence in these archaeal

Figure 7 Relative proportions of coding sequences with opposed transcriptional orientations in the archaeal genomic clones annotated in this study including all open reading frames (ORFs) (higher panel) or excluding the *spc* operon encoding the ribosomal proteins (lower panel). *Data derived from two contigs.

fosmids might indeed indicate that CO₂ fixation is possible in these organisms.

Multiple membrane transporters were tentatively identified on the basis of transmembrane segment predictions (Supplementary Table S1), but due to low similarity to any other transporters in the database, it is difficult to predict the kind of molecules that they might incorporate into the archaeal cells. A urea transporter was identified using BLASTX at the 5' end of KM3-86-C1, which had the highest similarity to a homologous sequence in *C. symbiosum* A (80% amino-acid similarity) followed by several eukaryotic hits, the closest of which came from the picoplanktonic alga *Ostreococcus tauri* (66% amino-acid similarity). However, neither *N. maritimus* SCM1 nor other sequenced archaea or archaeal fragments possessed a homologue of this gene. From the rest of transporter genes, only two, found in fosmids SAT1000-49-D2, SAT1000-23-F7 and AD1000-207-H3, could be clearly assigned to oligopeptide or dipeptide permease proteins (OppC and DppB types) (Supplementary Table S1). The potential for incorporating

oligopeptides in the cell, together with the possession of carbon fixation genes, suggests that, similarly to *C. symbiosum* A (Hallam *et al.*, 2006a,b), these meso- and bathypelagic crenarchaeota may behave as mixotrophs.

Genome fragments from Group 1A Crenarchaeota

The rest of our crenarchaeotal clones, namely, KM3-153-F8, AD1000-325-A12 and AD1000-23-H12, fell apart in the recently defined Crenarchaeota Group 1A. The clones studied here exhibited 16S rDNA identity levels below 97%, representing quite distinct OTUs, particularly AD1000-23-H12, with less than 86% 16S rDNA identity with the two others (Figure 2). Group 1A appears scarce in open oceanic libraries (DeLong *et al.*, 2006; Figure 1). However, in a previous study of the Ionian sampling site (Zaballos *et al.*, 2006) several 16S rRNA sequences (clusters Arch 1, 2, 3 and 4) that can be ascribed to the 1A group were detected, although none was found in samples of similar depth from the Greenland Sea, indicating that this group might be more

abundant in the bathypelagic Mediterranean. Along these lines, we identified and sequenced three clones in our Mediterranean libraries, but failed to detect any in the South Atlantic or Polar Front samples. This is the first report of genomic sequences from members of this group. Their average GC content was very low, 33.8, 31.5% if we exclude 16S rDNA (having 52.37–53.98% GC) (Table 2 and Supplementary Table S1). This value approached the average GC content in the rest of free-living planktonic Group I Crenarchaeota (35.9%), including the sequence of *N. maritimus* SCM1 (33.9%).

In accordance with their distant phylogenetic positions, there was no synteny between AD1000-23-H12 and the two other clones beyond the *rrn* operon (Figure 4 and Supplementary Table S1). Alternatively, since the number of ribosomal operons of members of the Group 1A is not known, we cannot rule out that the region sequenced belongs to a different ribosomal RNA operon. Also, as expected, the genes identified in all these three genomic fragments were totally different from those found in Group I Crenarchaeota. The lack of other Group 1A crenarchaeotal sequences in databases may partially explain why only 52.5% of the best BLAST hits were crenarchaeotal proteins. The remaining best BLAST hits were bacterial (31.2%) and euryarchaeotal (16.2%) proteins, a situation already observed during the analysis of the Group I clone DeepAnt-39 that was suggestive of high horizontal gene transfer rates (López-García *et al.*, 2004) (Supplementary Table S1).

The clones that were phylogenetically closer, AD1000-325-A12 and KM3-153-F8 (93.9% 16S rRNA identity) exhibited good synteny in their relatively small overlapping region (Figure 4). However, mean gene similarity was relatively low (43.7% nucleotide identity and 65.45% amino-acid similarity over the 8356 bp region that could be aligned excluding the *rrn* operon). This shared region upstream of the *rrn* operon contained a cluster of genes involved in the synthesis of aromatic amino acids. Downstream of it, both fosmids bore a luciferase-like protein, the next two genes being different in the two fosmids. Upstream of its overlapping region with AD1000-325-A12, KM3-153-F8 also contained a cluster of genes implicated in the biosynthesis of purine nucleotides. The more distant clone AD1000-23-H12 carried a set of genes that could be involved in nitrogen fixation. Immediately downstream of the *rrn* operon were genes for a putative molybdate ABC transporter (NifC-like ABC-type porter) followed by a homologous ORF similar to a dinitrogenase iron–molybdenum cofactor biosynthesis gene in *Clostridium phytofermentans* ISDg (*E*-value $10e-64$), which is required for nitrogen fixation in this organism (Mehta *et al.*, 2005; Leigh and Dodsworth, 2007). The remaining genes of this clone are mostly housekeeping and gave no clue as to possible metabolic capabilities that could be hypothesized for these archaea.

Organization of genomic fragments from Group II Euryarchaeota

Group II Euryarchaeota are reliably found in 16S rDNA and metagenomic surveys, and were present to a varying extent in our metagenomic libraries from different oceanic locations (Figure 1). To date, only one mesopelagic genomic sequence was available for this group, the 500-m deep planktonic clone DeepAnt-JYKC7 from the Antarctic Polar Front, which included one 16S rRNA gene and a ribosomal protein cluster (spectinomycin or *spc* operon) (Moreira *et al.*, 2004) (Table 1). In addition, a few other genomic fragments for this group were available at the time of writing this work, all of them from the photic region. They were the 23S rDNA-containing BAC clone 37F11 (Beja *et al.*, 2000b), the fosmid EF100-57A08 (Rich *et al.*, 2007) and three fosmids containing proteorhodopsin-like genes adjacent to the 16S rDNA, of which H70-59C08, contained also the *spc* operon (Frigaard *et al.*, 2006). We sequenced seven genomic clones from different deep-sea marine libraries, five of which were distributed in two different pan-oceanic OTUs (OTU E including AD1000-18-D2, KM3-130-D10 and SAT1000-15-B12 and OTU F including KM3-72-G3 and DeepAnt-15E7) according to the previously applied criterion of a 16S rRNA identity $\geq 98\%$ (Figure 3). From the remaining clones, KM3-85-F5 was quite close to these two OTUs, all the sequences forming with them and other environmental sequences a cluster with $\geq 96\%$ identity at the 16S rRNA level. This larger cluster was also relatively close to a group of sequences that contained the previously sequenced DeepAnt-JYKC7 (Figure 3). KM3-136-D10 was, in contrast, very distantly related to that cluster, having a significantly lower GC content, 45.31% versus an average of 58.7% for the remaining Group II euryarchaeotal clones (Table 2). Accordingly, the analysis of the codon usage and amino-acid composition for DeepAnt-15E7 and KM3-130-D10, both having high GC content, showed that triplets possessing G or C at second and third position were favoured for several amino acids, especially for abundant ones such as Gly (GCC), Arg (CGC), Ala (GCC), Val (GTC), Ser (AGC), Leu (CTG/CTG), Glu (GAG) and Lys (AAG) (Supplementary Figure S3). All fosmids from Group II (as well as those from Group III, see below) euryarchaeota had the 16S rRNA and 23S rRNA genes separated in the genome (Figure 8 and Supplementary Table S1), a general feature observed also for the Thermoplasmatales (Ruepp *et al.*, 2000).

Comparing to the situation within Group I Crenarchaeota, the most striking observation was the high degree of synteny shown by the different euryarchaeotal clones for similar (within OTUs) but also disparate (among distant OTUs) 16S rRNA genetic distances (compare Figures 4 and 8). Part of this synteny was due to the presence of the *spc* operon encoding ribosomal proteins and other ribosome-related proteins (translation initiation

Figure 8 Comparative genomic organization of fosmids from euryarchaeota. A neighbour-joining tree relating the different 16S rRNA markers found in the respective fosmids is shown on the left as reference. Conserved genomic regions between fosmids are indicated by grey shaded areas, grey intensity being a function of sequence similarity by BLASTN. Open reading frames (ORFs) are numbered from left (ORF 1 in each case) to right as shown in Supplementary Table S1. Particular ORFs mentioned in the text are highlighted by a graphic code.

factor SUI1 and protein translocase SecY), though not only. Thus, synteny was particularly well maintained further upstream of the *spc* operon whenever this region was sequenced, including DeepAnt-JYKC7 (Moreira *et al.*, 2004) and the clone KM3-136-D10, which was very divergent in terms of 16S rRNA sequence (88.7% identical to that of KM3-130-D10) (Figure 3) but also included genes encoding an ATP-dependent RNA helicase, the B subunit of a heterodisulphite reductase, a characteristic protein with three fused domains, Hsp60 and a conserved transmembrane protein (Figure 8 and Supplementary Table S1). As done for the crenarchaeota, and since co-linearity was generally well conserved, we applied current programs for the detection of recombination breakpoints, but we failed to detect any with confidence (data not shown). The genomic stability and apparent absence of intra-genomic recombination events leading to rearrangement observed in the surroundings of the 16S rRNA gene contrasts markedly with the active intra- and inter-genomic recombination and gene shuffling occurring in Group I Crenarchaeota around

the *rrn* operon. This difference is further reinforced by the transcriptional orientation of genes in this area relative to that of the 16S rRNA gene, as approximately 90% of the genes had the same transcriptional sense as the 16S rRNA gene. This trend for genes to be oriented in the same direction was also clearly visible even when we removed the *spc* operon cluster from our analysis (Figure 7).

Though overall synteny was conserved, the genetic distance at protein markers was noticeable even between closely related OTUs. For instance, sequence divergence at the *spc* cluster from two clones picked up as examples from the closely related OTUs E (KM3-130-D10) and F (DeepAnt-15E7) was somewhat higher than expected for such highly conserved informational proteins. The average nucleotide and amino-acid identities were only 78.29 and 85.65%, respectively. As a reference, we compared the sequence divergence for *spc* operon genes in three different *Methanosarcina* species, *M. mazei* Go1, *M. barkeri* strain Fusaro and *M. acetivorans* C2A (with a 16S rRNA identity ca. 97%). Nucleotide identity values were 84.68% (Go1

versus Fusaro) and 91.5% (Fusaro versus C2A), in parallel with 89.85 and 93.66% amino-acid similarity, respectively. The relatively low similarity values observed for Group II were not due to potential gene exchange by horizontal gene transfer of ribosomal protein genes from foreign donors, since phylogenetic trees for the concatenated sequences as well as all individual trees were congruent with the phylogenetic relationships shown by the 16S rRNA (data not shown). They appear to be simply explained as the consequence of high mutational activity. Even if most ribosomal protein genes had a very small ratio of non-synonymous versus synonymous substitutions (dN/dS) showing strong purifying selection, in some cases we obtained evidence suggesting that positive selection may be occurring. For example, the ribosomal protein S3 in DeepAnt-15E7 (OTU F) and KM3-130-D10/SAT1000-15-B12 (OTU E) gave dN/dS values of 1.357 and 1.355, respectively. Whether this is the consequence of actual positive selection in response to particular selective pressures, for instance antibiotic resistance, or other causes remains speculative at this point. At any rate, our observations would be consistent with mutation playing a more prominent role than intra-genomic, and perhaps also inter-genomic, recombination in the genome evolution of Group II Euryarchaeota. In fact, although a high level of homologous recombination between members of the same or related OTUs could explain the high level of synteny conservation, this would be at odds with the apparent high accumulation of mutations observed that should be neutralized under the effect of intense homologous recombination.

Metabolic clues from Group II Euryarchaeota genome fragments

In agreement with their meso- and bathypelagic provenance, none of our Group II Euryarchaeota genomic clones possessed the proteorhodopsin-like gene that was found immediately downstream of the 16S rRNA gene in some euryarchaeota from the photic region (Frigaard *et al.*, 2006). Apart from the possible phototrophy of those surface euryarchaeota after the horizontal acquisition of a proteorhodopsin-like light-sensitive proton pump, the only hints about possible metabolic capabilities for marine Group II Euryarchaeota were those pointed out by the previous analysis of clone DeepAnt-JYKC7. It harboured a cluster of two protein genes that might be co-transcribed encoding the putative components of an anaerobic respiratory chain (heterodisulphide reductase subunit B and a multidomain protein with a ferredoxin, a flavodoxin and a succinate dehydrogenase/fumarate reductase Fe-S subunit domain) (Moreira *et al.*, 2004). As mentioned above, these genes were also found in our five fosmids overlapping this region (Figure 8 and Supplementary Table S1).

In addition, some genes indicating potential metabolic capabilities were detected downstream of the 16S rRNA genes in the clones that covered that region, that is, DeepAnt-15E7 and KM3-72-G3 (OTU F), as well as KM3-130-D10 and AD1000-18-D2 (OTU E). They contained a glutamine ABC transport system with the periplasmic subunit, two permeases, and an ATP-binding protein. At the further downstream end of fosmid AD1000-18-D2, two genes coding for the carbamoyl-phosphate synthase complex, involved in both the urea cycle and the biosynthesis of arginine and pyrimidines, were identified. They were found also in the surface clones EF100-57A08 (Rich *et al.*, 2007), HF70-39H11 (Frigaard *et al.*, 2006) and the river sediment clone GZfos14B8 belonging to anaerobic methanotrophic archaea (Hallam *et al.*, 2004). Although the region downstream of the 16S rRNA gene was syntenic in our clones, only a few of these genes were conserved in the surface clone H70-59C08 (Rich *et al.*, 2007), an ATP-NAD kinase protein, a membrane protein flotillin-like and a 3-ketoacyl-CoA thiolase. Perhaps the most interesting observation in this region downstream of the 16S rRNA gene came from the identification of a molybdopterin oxidoreductase, a 4Fe-4S ferredoxin (iron-binding protein) and a TraB/PrgY-like protein. Interestingly, the *dsm* operon described for *Halobacterium* sp. NRC1 involved in the anaerobic respiration of dimethylsulphoxide (DMSO) (Muller and DasSarma, 2005) has a similar structure. Although not all the genes were the same, equivalent domains were always found. This might indicate that these euryarchaeota use DMSO as an electron acceptor.

In summary, these Group II euryarchaeotal fragments contained not only ribosomal and other housekeeping genes but seemed enriched in putative anaerobic respiration components, which might suggest that they gain energy by anaerobically respiring various substrates.

Genome fragments from Group III Euryarchaeota

Euryarchaeota Group III was first identified from environmental 16S rRNA gene libraries by Fuhrman and Davis (1997) in samples retrieved from the Northeast Pacific at 500 and 300 m depth, and was later detected at 941 m depth in the Alboran Sea (Mediterranean) (Massana *et al.*, 2000) and at 3000 m depth in the Antarctic Polar Front (López-García *et al.*, 2001), suggesting that they were specific inhabitants of the deep ocean. Group III members were also detected in metagenomic libraries throughout the ALOHA water column in the Pacific, although at lower frequencies in surface waters (DeLong *et al.*, 2006). Though they are generally little represented in environmental studies, they appeared more relatively abundant in our SAT1000 library (Figure 1). We sequenced three fosmid clones from three different libraries corresponding to three geographic locations. Two of them, AD1000-40-D7

and SAT1000-53-B3 belonged to a same pan-oceanic OTU (99.1% 16S rDNA identity, OTU D) (Figure 3), had similar low GC content (Table 2) and were largely syntenic (Figure 8 and Supplementary Table S1). Despite their relatedness at the 16S rDNA locus, the similarity of the overlapping ORFs at amino-acid level was only 88.88%. The more distant clone KM3-28-E8 (92.7% identity to AD1000-40-D7) had higher GC content (54.8%) and had only one gene in common with the two other fosmids, the first truncated ORF, coding for a hypothetical protein bearing only 31% similarity to its homologue in the other clones. The information about genome organization and dynamics in this group of archaea is very limited, and very little can be said from comparing only these three fosmids. Nevertheless, since only approximately 50% of the ORFs had the same transcriptional orientation as the 16S rRNA gene (Figure 7), a higher rearrangement frequency could be hypothesized for these archaea than that observed for the Group II in the same genomic area.

As in the case of Group II archaea, a significant number of genes in SAT1000-53-B3 and AD1000-40-D7 were housekeeping. Many of them coded for components of the translation apparatus, such as a valine tRNA, an alanyl-tRNA synthetase, an aspartyl-tRNA synthetase, a putative metal-dependent hydrolase related to alanyl-tRNA synthetase, the signal recognition particle protein Srp54 and the translation termination factor aRF1 (Supplementary Table S1). Others were related to DNA repair (for example, an endonuclease IV homologue) or transcription (a putative σ^{54} -related protein). A peptidyl-prolyl *cis-trans* isomerase (*ppiD*), not previously described in archaea, was also identified. PPIase is an enzyme that accelerates protein folding by catalysing the *cis-trans* isomerization of proline imidic peptide bonds in oligopeptides (Dartigalongue and Raina, 1998). Two genes (ORF 11 and 17 in SAT1000-53-B3, and ORF 18 and 24 in AD1000-40-D7) could be involved in the recently described and unusual wyeosine post-transcriptional modification of tRNA. The wyeosine family of tricyclic ribonucleosides is one of the more complex and structurally interesting post-transcriptional modifications in tRNAs, and has so far only been found in the archaeal and eukaryal domains (Zhou *et al.*, 2004). The first ORF was homologous to a TYW3 domain protein, which is usually involved in wybutosine (yW) biosynthesis in yeast, and the second ORF was homologous to a SAM-dependent methyltransferase related to tRNA (uracil-5-)-methyltransferase modifications (de Crecy-Lagard, 2007).

Regarding potential metabolic clues, one of the most significant proteins found in AD1000-40-D7 and SAT1000-53-B3 was a homologue of an ammonium permease, AmtB, with an associated histidine kinase domain. Proteins with similar structure have been identified so far only in the bacterial species *Parvibaculum lavamentivorans* DS-1 (65% similar-

ity), *Planctomyces maris* DSM 8797 (62%) and in the anaerobic ammonium oxidizer *Candidatus Kueenia stuttgartiensis* (60%). Adjacent to this gene in AD1000-40-D7, there was a duplication of the histidine kinase domain present in this putative ammonium transporter. Next to it was a gene homologous to CheY, a chemotaxis response regulator. This two-component system combination is evocative of an ammonium sensor and transporter resembling those found in ammonia-oxidizing bacteria (Arp *et al.*, 2007). Fosmid SAT1000-53-B3 also contained genes for the carbamoyl-phosphate synthase complex, as in the Euryarchaeota Group II (see above). The more distant clone KM3-28-E8 harboured a gene potentially involved in anaerobic respiration (Acyl-CoA dehydrogenase short-chain-specific gene), as well as three putative genes that might be involved in C fixation by the 3-hydroxipropionate pathway or one variant of it. These were 3-hydroxyacyl-CoA dehydrogenase (*fadB*), malonyl-CoA synthetase (*fadD*) and enoyl-CoA hydratase (*fadB*) (Supplementary Table S1). Whether these enzymes are specifically involved in autotrophic C fixation or in other pathways of central metabolism, for example, fatty acid synthesis, remains uncertain.

Discussion

Marine pelagic archaea are among the most widespread and abundant marine microbes, having been found in many marine environmental 16S rRNA gene libraries since they were first discovered (DeLong, 1992; Fuhrman *et al.*, 1992). Despite their obvious interest, relatively little is known about their ecology and evolution. Fortunately, the accumulation of molecular, genomic and metagenomic data combined with environmental biochemistry and the first cultivation success has led to a fundamental breakthrough concerning the ecological role of the most easily encountered marine archaea, the Group I Crenarchaeota, which appear to be aerobic nitrifiers (Konneke *et al.*, 2005; Hallam *et al.*, 2006a,b; Wuchter *et al.*, 2006; Lam *et al.*, 2007). However, this type of information is missing for most of the other planktonic archaea that also appear to be widespread in oceans, and nothing is known in general about archaeal population dynamics and evolution in various oceanic regions. To gain additional information about marine pelagic archaea, and since the vast majority of them remain uncultivable, we detected, sequenced and comparatively analysed genomic clones that possessed archaeal 16S rRNA genes from four meso- and bathypelagic planktonic metagenomic libraries representing different oceanic regions and water masses (Table 1). We selected and analysed both clones belonging to the divergent and less abundant Group 1A Crenarchaeota and Group III Euryarchaeota, for which nearly nothing is known, and a

number of very closely related clones forming defined OTUs within the Group I Crenarchaeota and Group II Euryarchaeota. Our study targeted three objectives. First, metagenomic library screening would generate information about the relative distribution of different archaeal groups in various deep-sea regions. In addition, sequencing genomic clones carrying a reference phylogenetic marker would eventually lead to (i) the identification of metabolic genes and hypothesizing functions for given archaeal lineages and (ii) the characterization of genome structure in comparable genomic surroundings, which might elucidate particular lineage-specific genome dynamics and lead to hypotheses about evolutionary trends.

It is commonly accepted that Group I Crenarchaeota are more abundant in oligotrophic, deep waters, whereas Group II Euryarchaeota are either more abundant in surface waters or equally rare throughout the water column (Karner *et al.*, 2001; Herndl *et al.*, 2005), something that seemed to be confirmed by some metagenomic studies (DeLong *et al.*, 2006; Martin-Cuadrado *et al.*, 2007). However, extensive comparative deep-sea studies to fully appreciate the distribution and potential phylogeography of marine pelagic archaea are lacking and, therefore, drawing general conclusions from few locations and collection dates may be unreliable. In this sense, our screening of four metagenomic libraries from the Antarctic Polar Front, the South Atlantic, and the Mediterranean Ionian and Adriatic seas, containing from ca. 5500 to 39000 clones, revealed important different relative proportions of Groups I and 1A Crenarchaeota and Groups II and III Euryarchaeota in the different libraries (Figure 1). Since all potential cloning biases were similar, and since at least Groups I and II archaea carry likely one single 16S rRNA gene (which can be deduced from their comparative genome organization), these differences could reflect actual relative frequencies in the environment. Many more studies on spatial and temporal relative abundances of marine pelagic archaea will be needed to understand their population dynamics. Despite differences in relative abundances, we did not observe any clear phylogeographical pattern. In contrast, for the more abundant archaeal Groups I (Crenarchaeota) and II (Euryarchaeota), we identified clones belonging to the same OTUs (that is, having $\geq 98\%$ identical 16S rRNA genes) from geographically very distant metagenomic libraries (Mediterranean and South Atlantic/Antarctic Polar Front). Several of these OTUs also include environmental sequences from Pacific 16S rRNA gene and/or metagenomic libraries (Figures 2 and 3). This suggests that at least some of what might actually be archaeal species have a true pan-oceanic distribution, though their relative abundances may vary spatially and temporally. Stratification in the water column possibly determines the strongest spatial pattern. Indeed, most of our clones clustered with deep-sea sequences, and

we were unable to detect putatively phototrophic euryarchaeota (Frigaard *et al.*, 2006) in our metagenomic libraries.

The protein genes identified in the 22 archaeal genomic fragments sequenced corresponded in large part to housekeeping machinery (translation, transcription, DNA repair, etc.) and to some conserved hypothetical proteins. However, a number of genes pointed out to some clues indicating potential metabolic capabilities in different archaeal lineages. We did not detect any *amoA* gene in Group I Crenarchaeota that might confirm the suspicion that they are ammonia oxidizers, as is the case of *N. maritimus* (Konneke *et al.*, 2005), *C. symbiosum* (Hallam *et al.*, 2006a) and some Group I soil fosmids (Treusch *et al.*, 2005). Yet, we detected various genes, having homologues in *C. symbiosum* and *N. maritimus*, involved in 3-hydroxypropionate-based pathways of CO₂ fixation, suggesting that these organisms can also fix carbon autotrophically. Interestingly, genes encoding oligopeptide transporters were found in clones from two Group I archaeal OTUs, suggesting that at least some Group I archaeal species could be mixotrophic, as is the case of *C. symbiosum* (Hallam *et al.*, 2006a, b). Mixotrophy is being increasingly detected in marine microbes. It includes not only phototrophic bacteria (Beja *et al.*, 2000a) and euryarchaeota (Frigaard *et al.*, 2006) but also lithoheterotrophic microbes able to gain extra energy from CO oxidation in surface (Moran and Miller, 2007) and deep-sea water (Martin-Cuadrado *et al.*, 2007). In a recent work, amplified and sequenced *amoA* genes at different depths in the North Pacific Subtropical Gyre showed a strong phylogenetic partitioning with depth. This was correlated with quantitative distributions of Group 1A sequences and was interpreted as an indication of Group 1A being putative ammonia oxidizers as well (Mincer *et al.*, 2007). We did not identify any *amoA* gene in the three clones that we sequenced from this archaeal group. Nonetheless, we found genes involved in nitrogen fixation in one of them. Genes for nitrogen fixation have also been found in some hyperthermophilic methanogens (Mehta and Baross, 2006) and other deep sea so far uncultivated archaea (Mehta *et al.*, 2005). At any rate, this observation brings to light the possibility that marine pelagic crenarchaeota are not only key but also multifaceted players in the nitrogen cycle.

Furthermore, members of the Group III Euryarchaeota might also utilize the oxidation of ammonia as a means to gain energy, as we detected genes for a histidine kinase two-component system that resembled that found in ammonia oxidizing bacteria, including marine planktonic planctomycetes and the anammox planctomycete *K. stuttgartiensis*. Whether Group III Euryarchaeota can actually oxidize ammonia and whether they do it aerobically or anaerobically should deserve future investigations. Finally, regarding the potential energy metabolism of marine Group II Euryarchaeota, we

systematically observed the presence of genes possibly involved in an anaerobic respiration chain that had already been described in the Antarctic Polar Front clone DeepAnt-JYKC7 (Moreira *et al.*, 2004). In addition to this, the presence of a putative DMSO respiration cluster suggested that these organisms could use DMSO as electron acceptor. Although the significance of DMSO respiration in the global turnover of marine organo-sulphur compounds derived from algal remains needs to be established, the fact that several marine bacteria including members of the genera *Shewanella* and *Rhodobacter* carry a *dms* operon (McCrindle *et al.*, 2005), as well as halophilic and perhaps Group II Euryarchaeota, might point to a significant role of this form of anaerobic respiration for the marine sulphur cycle.

Metagenomic analyses are beginning to make possible the study of population genomics in the environment. One of the most interesting observations of our work comes from the realization that differential genomic dynamics appear to be operating in different archaeal phylogenetic groups. This was particularly evident between Group I Crenarchaeota and Group II Euryarchaeota, for which we had the largest sampling. Thus, whereas gene shuffling by intra-genomic recombination was very important around the *rrn* operon within and among OTUs in Group I Crenarchaeota (Figure 4), co-linearity was strikingly maintained even across relatively distant OTUs within Group II Euryarchaeota (Figure 8). This opposite trend for equivalent genomic regions around the likely single-copy 16S rRNA gene in the capacity to rearrange was further supported by the relative orientation of genes in these genome fragments. Whereas approximately 90% of the genes in Group II genome fragments had the same transcriptional orientation as the 16S rRNA gene, only half of the genes in Group I clones did (Figure 7). Likewise, inter-genomic recombination leading to both genetic exchange within populations of the same or related OTUs (via homologous recombination, for example, at the *rrn* operon and the EF-2 gene; Supplementary Figures S2A and Figure 6) and with far more distant organisms resulting in horizontal gene transfer (López-García *et al.*, 2004), appeared important in Group I Crenarchaeota. By contrast, we failed to detect with confidence potential recombination breakpoints in Group II Euryarchaeota fosmids. This and the fact that sequence divergence by mutation is relatively high in these organisms tend to support lower inter-genomic homologous recombination rates in Group II Euryarchaeota, which would otherwise counter-balance mutational activity, as compared to Group I Crenarchaeota. This indicates that intra- and inter-genomic recombination is an active process shaping genome evolution in Group I members. By contrast, sequence divergence in protein genes relative to 16S rRNA genetic distances appears to be much higher in Group II Euryarchaeota than in Group I Crenarch-

aeota. Moreover, evidence of positive selection was observed in one gene coding for the ribosomal protein S3. Therefore, mutation fixation might be the predominant mechanism in the evolution of Group II euryarchaeotal genomes. Of course, these observations are restricted so far to a small genomic region around the apparently single 16S rRNA gene in the genome, which we are using here as a proxy to hypothesize similar trends at the whole genome scale. This hypothesis may be supported by the general observation that the *rrn* operon surroundings are normally regions harbouring important housekeeping genes that are highly expressed and, therefore, subjected to higher selection than other genomic regions such as the terminus of replication. Nonetheless, falsifying this hypothesis will require full genome comparison of these archaea in the future.

Interestingly, excluding the truncated genomic clone KM3-136-D10, all Group II Euryarchaeota genome fragments had high GC content (average 58.7%), whereas all Group I Crenarchaeota had low GC content (35.9%) (Table 2). Differences in GC content in different phylogenetic clades might be simply the consequence of phylogenetic history of particular lineages. Yet they might be as well indicators of particular lifestyle strategies and genomic constraints. For instance, intracellular parasites tend to possess AT-rich genomes, since there is a mutational bias towards AT derived from the relaxation of selective pressures acting on some gene categories, such as DNA repair genes (Rocha and Danchin, 2002; Moran, 2003), while organisms living in environments where competition is strong and selective pressures high would tend to maintain higher GC genomes. According to this view, there could be two alternative hypotheses to explain the differences in GC content observed. First, it could be speculated that marine Group I Crenarchaeota are actually free-living planktoners (*C. symbiosum*, whose genome is GC rich, is a sponge symbiont and would have to compete with the sponge microbiota), whereas Group II Euryarchaeota might rather live associated with sinking particles and be therefore subjected to strong competition and selection. Though we pre-filtered samples by 5 µm-pore size filters, it might be possible that some particle-associated microbiota was recovered. Indeed, the GC values of the total sequence determined for different size ranges in the Sargasso shotgun metagenomic database decreases along with the pore size of the filter used, although there might be some potential bias introduced by the capture of eukaryotic cells using larger pore sizes (Supplementary Figure S4). This hypothesis could be eventually supported by the presence in the Group II Euryarchaeota of genes involved in anaerobic respiration (more likely to occur within particulate matter) or resembling those of planctomycetes, which are thought to be physically associated with particles (Woebken *et al.*, 2007).

However, this first hypothesis would be at odds with fluorescent *in situ* hybridization studies showing that both marine pelagic crenarchaeota and euryarchaeota are free-living and not associated with particles (Woebken *et al.*, 2007). If this is indeed the case and both archaeal groups were actually planktonic and free-living, a second, alternative hypothesis to explain the observed differences in genomic GC content could be that Group I Crenarchaeota had shorter generation times, behaving more like r-strategists or 'bloomers', as compared to Group II Euryarchaeota, which would have longer generation times relative to Group I Crenarchaeota, sustaining growth when resources are low and behaving as K-strategists or 'survivalists'. These different lifestyle strategies are indeed related to the N/P ratios in the cell in marine pelagic ecosystems (Klausmeier *et al.*, 2004; Arrigo, 2005). Having low GC genomes is a way to save N, a limiting element in oligotrophic oceanic waters (AT having seven N atoms, one N atom less than GC pairs, making a significant difference at the whole genome scale). Therefore, pelagic microbes that grow faster and reach larger population sizes would tend to accumulate low GC genomes not as a consequence of selective pressure relaxation (as occurs in parasites) but, in contrast, because there would be a strong selective pressure for r-strategists to save N under environmental conditions where this resource is limiting. In fact, this trend has been observed in free-living marine *Prochlorococcus* species genomes. These organisms undergo a strong purifying selection, so that low GC genome content has been interpreted as a strategy to save energy and nutrients (Dufresne *et al.*, 2005). This tendency to have lower GC genomes as a consequence of strong purifying selection under nutrient limitation would be particularly favoured in Group I Crenarchaeota as, being ammonia oxidizers, they would tend to deviate most of the available reduced nitrogen (ammonia) to dissimilatory energy metabolism rather than to assimilatory pathways in a globally N-limiting oligotrophic ocean.

In conclusion, our comparative genomic analyses reveal remarkably different genomic dynamics in Groups I and II marine archaea that could be linked to opposite life strategies, both being subject to strong selection, in these two groups of organisms. In Group I Crenarchaeota, the possession of low GC genomes in potentially faster growing microbes (as compared to Group II Euryarchaeota) might be correlated with higher recombination levels. In fact, recombination appears to be higher in AT-rich regions in chromosomes (terminus of replication, insertion elements). In planktonic Group II Euryarchaeota, strong selection pressures endured by survivalist strategists would favour GC-rich genomes and low recombination frequencies. These hypotheses remain open to be tested in the future.

Acknowledgements

We thank M Zaballos and L Monticelli for the collection of samples from the Adriatic Sea. This study was supported by the collaborative Spanish–French Picasso project HF 2005-0149 (European Integrated Action Program), by an ATIP Plus grant of the French Centre National de la Recherche Scientifique (CNRS) to PLG and by the Spanish Ministerio de Educación y Ciencia (MEC) Grant CTM2005-04564/MAR to FRV. Fosmid sequencing was possible thanks to the Gordon and Betty Moore Foundation, USA. A-BM-C was supported by MEC and FEMS fellowships.

References

- Altschul SF, Madden TL, Schaffer AA, 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.
- Arp DJ, Chain PS, Klotz MG. (2007). The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. *Annu Rev Microbiol* **61**: 503–528.
- Arrigo KR. (2005). Marine microorganisms and global nutrient cycles. *Nature* **437**: 349–355.
- Bano N, Ruffin S, Ransom B, Hollibaugh JT. (2004). Phylogenetic composition of Arctic Ocean archaeal assemblages and comparison with Antarctic assemblages. *Appl Environ Microbiol* **70**: 781–789.
- Barns SM, Delwiche CF, Palmer JD, Pace NR. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci USA* **93**: 9188–9193.
- Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al.* (2000a). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902–1906.
- Beja O, Koonin EV, Aravind L, Taylor LT, Seitz H, Stein JL *et al.* (2002a). Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microbiol* **68**: 335–345.
- Beja O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T *et al.* (2002b). Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630–633.
- Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP *et al.* (2000b). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516–529.
- Berg IA, Kockelkorn D, Buckel W, Fuchs G. (2007). A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. *Science* **318**: 1782–1786.
- Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics* **21**: 3422–3423.
- Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.
- Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG *et al.* (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**: 3497–3500.
- Darling AC, Mau B, Blattner FR, Perna NT. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**: 1394–1403.

- Dartigalongue C, Raina S. (1998). A new heat-shock gene, *ppiD*, encodes a peptidyl-prolyl isomerase required for folding of outer membrane proteins in *Escherichia coli*. *EMBO J* **17**: 3968–3980.
- de Crecy-Lagard V. (2007). Identification of genes encoding tRNA modification enzymes by comparative genomics. *Methods Enzymol* **425**: 153–183.
- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–4641.
- DeLong EF. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89**: 5685–5689.
- DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.
- Dufresne A, Garczarek L, Partensky F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6**: R14.
- Edgar RC. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.
- Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci USA* **102**: 14683–14688.
- Frigaard NU, Martinez A, Mincer TJ, DeLong EF. (2006). Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847–850.
- Fuhrman JA, Davis AA. (1997). Widespread Archaea and novel Bacteria from the deep sea as shown by 16S rRNA gene sequences. *Mar Ecol Prog Ser* **150**: 275–285.
- Fuhrman JA, McCallum K, Davis AA. (1992). Novel major archaeobacterial group from marine plankton. *Nature* **356**: 148–149.
- Garcia-Martinez J, Rodriguez-Valera F. (2000). Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* **9**: 935–948.
- Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J et al. (2006a). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci USA* **103**: 18296–18301.
- Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM et al. (2006b). Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* **4**: e95.
- Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM et al. (2004). Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **305**: 1457–1462.
- Herndl GJ, Reinthaler T, Teira E, van Aken H, Veth C, Pernthaler A et al. (2005). Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. *Appl Environ Microbiol* **71**: 2303–2309.
- Huber JA, Johnson HP, Butterfield DA, Baross JA. (2006). Microbial life in ridge flank crustal fluids. *Environ Microbiol* **8**: 88–99.
- Ingalls AE, Shah SR, Hansman RL, Aluwihare LI, Santos GM, Druffel ER et al. (2006). Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* **103**: 6442–6447.
- Jobb G, von Haeseler A, Strimmer K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**: 18.
- Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.
- Klausmeier CA, Litchman E, Daufresne T, Levin SA. (2004). Optimal nitrogen-to-phosphorus stoichiometry of phytoplankton. *Nature* **429**: 171–174.
- Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.
- Konstantinidis KT, Tiedje JM. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**: 2567–2572.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.
- Kumar S, Tamura K, Nei M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**: 150–163.
- Lam P, Jensen MM, Lavik G, McGinnis DF, Muller B, Schubert CJ et al. (2007). Linking crenarchaeal and bacterial nitrification to anammox in the Black Sea. *Proc Natl Acad Sci USA* **104**: 7104–7109.
- Leigh JA, Dodsworth JA. (2007). Nitrogen regulation in bacteria and archaea. *Annu Rev Microbiol* **61**: 349–377.
- López-García P, Brochier C, Moreira D, Rodríguez-Valera F. (2004). Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ Microbiol* **6**: 19–34.
- López-García P, Moreira D, López-López A, Rodríguez-Valera F. (2001). A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environ Microbiol* **3**: 72–78.
- 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.
- Martin DP, Williamson C, Posada D. (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**: 260–262.
- Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, Strittmatter A et al. (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2**: e914.
- Massana R, DeLong EF, Pedros-Alio C. (2000). A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Appl Environ Microbiol* **66**: 1777–1787.
- McCrindle SL, Kappler U, McEwan AG. (2005). Microbial dimethylsulfoxide and trimethylamine-N-oxide respiration. *Adv Microb Physiol* **50**: 147–198.
- Mehta MP, Baross JA. (2006). Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. *Science* **314**: 1783–1786.
- Mehta MP, Huber JA, Baross JA. (2005). Incidence of novel and potentially archaeal nitrogenase genes in the deep Northeast Pacific Ocean. *Environ Microbiol* **7**: 1525–1534.
- Mincer TJ, Church MJ, Taylor LT, Preston C, Karl DM, DeLong EF. (2007). Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environ Microbiol* **9**: 1162–1175.

- Moran MA, Miller WL. (2007). Resourceful heterotrophs make the most of light in the coastal ocean. *Nat Rev Microbiol* **5**: 792–800.
- Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6**: 512–518.
- Moreira D, Rodríguez-Valera F, López-García P. (2004). Analysis of a genome fragment of a deep-sea uncultivated Group II euryarchaeote containing 16S rDNA, a spectinomycin-like operon and several energy metabolism genes. *Environ Microbiol* **6**: 959–969.
- Moreira D, Rodríguez-Valera P, López-García P. (2006). Genome fragments from mesopelagic Antarctic waters reveal a novel deltaproteobacterial group related to the myxobacteria. *Microbiology* **152**: 505–517.
- Muller JA, DasSarma S. (2005). Genomic analysis of anaerobic respiration in the archaeon *Halobacterium* sp. strain NRC-1: dimethyl sulfoxide and trimethylamine *N*-oxide as terminal electron acceptors. *J Bacteriol* **187**: 1659–1667.
- Murray AE, Preston CM, Massana R, Taylor LT, Blakis A, Wu K *et al.* (1998). Seasonal and spatial variability of bacterial and archaeal assemblages in the coastal waters near Anvers Island, Antarctica. *Appl Environ Microbiol* **64**: 2585–2595.
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.
- Page RD. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**: 357–358.
- Philippe H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res* **21**: 5264–5272.
- Posada D, Crandall KA. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci USA* **98**: 13757–13762.
- Ray SC. (1998). SimPlot for Windows (version 1.6). Baltimore: distributed by the author.
- Reysenbach AL, Liu Y, Banta AB, Beveridge TJ, Kirshtein JD, Schouten S *et al.* (2006). A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature* **442**: 444–447.
- Rice P, Longden I, Bleasby A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.
- Rich VI, Konstantinidis K, Delong EF. (2007). Design and testing of ‘genome-proxy’ microarrays to profile marine microbial communities. *Environ Microbiol* **10**: 506–521.
- Rocha EP, Danchin A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**: 291–294.
- Rodionov DA, Mironov AA, Gelfand MS. (2002). Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res* **12**: 1507–1516.
- Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW *et al.* (2000). The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**: 508–513.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.
- Sawyer SA. (1999). GENECONV: a computer package for the statistical detection of gene conversion.
- Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. (1998). Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. *J Bacteriol* **180**: 5003–5009.
- Schleper C, Jurgens G, Jonscheit M. (2005). Genomic studies of uncultivated archaea. *Nat Rev Microbiol* **3**: 479–488.
- Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**: 591–599.
- Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS *et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**: 22–28.
- Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk HP, Schleper C. (2005). Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ Microbiol* **7**: 1985–1995.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.
- Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.
- Woebken D, Fuchs BM, Kuypers MM, Amann R. (2007). Potential interactions of particle-associated anammox bacteria with bacterial and archaeal partners in the Namibian upwelling system. *Appl Environ Microbiol* **73**: 4648–4657.
- Wuchter C, Abbas B, Coolen MJ, Herfort L, van Bleijswijk J, Timmers P *et al.* (2006). Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* **103**: 12317–12322.
- Zaballos M, Lopez-Lopez A, Ovreas L, Bartual SG, D’Auria G, Alba JC *et al.* (2006). Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. *FEMS Microbiol Ecol* **56**: 389–405.
- Zhou S, Sitaramaiah D, Noon KR, Guymon R, Hashizume T, McCloskey JA. (2004). Structures of two new ‘minimalist’ modified nucleosides from archaeal tRNA. *Bioorg Chem* **32**: 82–91.
- Zmasek CM, Eddy SR. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* **17**: 821–828.

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