Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group

David Moreira,¹ Francisco Rodríguez-Valera²
and Purificación López-García¹

Correspondence

Purificación López-García
puri.lopez@ese.u-psud.fr

¹Unité d'Ecologie, Systématique et Evolution, CNRS UMR 8079, Université Paris-Sud, 91405 Orsay Cedex, France

²División de Microbiología and Evolutionary Genomics Group, Universidad Miguel Hernández, Campus de San Juan, 03550 San Juan de Alicante, Spain

Phylogenetic screening of 3200 clones from a metagenomic library of Antarctic mesopelagic picoplankton allowed the identification of two bacterial 16S-rDNA-containing clones belonging to the *Deltaproteobacteria*, DeepAnt-1F12 and DeepAnt-32C6. These clones were very divergent, forming a monophyletic cluster with the environmental sequence GR-WP33-58 that branched at the base of the myxobacteria. Except for the possession of complete *rrn* operons without associated tRNA genes, DeepAnt-1F12 and DeepAnt-32C6 were very different in gene content and organization. Gene density was much higher in DeepAnt-32C6, whereas nearly one-third of DeepAnt-1F12 corresponded to intergenic regions. Many of the predicted genes encoded by these metagenomic clones were informational (i.e. involved in replication, transcription, translation and related processes). Despite this, a few putative cases of horizontal gene transfer were detected, including a transposase. DeepAnt-1F12 contained one putative gene encoding a long cysteine-rich protein, probably membrane-bound and Ca²⁺-binding, with only eukaryotic homologues. DeepAnt-32C6 carried some predicted genes involved in metabolic pathways that suggested this organism may be anaerobic and able to ferment and to degrade complex compounds extracellularly.

Received 8 June 2005

Revised 31 October 2005

Accepted 6 November 2005

INTRODUCTION

Marine microbiology has benefited greatly in the last 20 years from the development of molecular tools to analyse the genetic diversity of microbial communities. The first molecular surveys based on 16S rRNA gene amplification of oceanic picoplankton revealed the existence of novel groups of *Bacteria* and *Archaea* (DeLong, 1992; Fuhrman *et al.*, 1992; Giovannoni *et al.*, 1990), opening the door to the application of this strategy to other environments and to the recognition that microbial diversity on our planet was far larger than previously thought (Pace, 1997). The continuous exploration of microbial diversity in different oceanic regions using a variety of cultivation-independent approaches has allowed the identification of various groups of non-cultivable organisms, some of which appear abundant and are probably major players in nutrient cycling (DeLong, 2001; Karner *et al.*, 2001; Morris *et al.*, 2002; Rappé &

Giovannoni, 2003). However, the functional study of these micro-organisms is severely hampered by the lack of tools providing clues as to their physiological capabilities. Environmental genomics (or metagenomics) has recently been revealed as a powerful source of information about the gene content of non-cultivable organisms and, hence, by comparative genomics, a source of predicted metabolic activities that are testable. A paradigmatic example was the discovery of a novel type of phototrophy in a gammaproteobacterial lineage, SAR86, that is abundant and widespread in the photic zone (Beja *et al.*, 2000a, 2001). This proteorhodopsin-based phototrophy appears to be common in other photic bacterial lineages as well (De La Torre *et al.*, 2003; Venter *et al.*, 2004). Nevertheless, in contrast to such fortuitously clear-cut examples and the relatively straightforward reconstruction of genome scaffolds by massive shotgun sequencing of low-diversity environments (Hallam *et al.*, 2004; Tyson *et al.*, 2004), most environmental genomic data accumulated to date are difficult to interpret. In many cases, putative genes bear resemblance to genes of unknown function; in others, they belong to large protein families whose precise function in the organism is difficult to predict. However, the accumulation of environmental data of this kind is a stepping stone for future comparative genomic studies that will eventually yield comprehensive conclusions.

Abbreviations: HGT, horizontal gene transfer; ITS, intergenic spacer; MRG, myxobacteria-related group.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are DQ267495 (DeepAnt-1F12) and DQ267496 (DeepAnt-32C6).

Supplementary figures are available with the online version of this paper.

Compared to the photic oceanic layers, the deep sea remains largely unexplored. Most efforts to describe its microbial diversity have been devoted to the study of deep-sea hydrothermal vents and, to a lesser extent, cold seep areas and deep-sea sediments (Inagaki *et al.*, 2002; Jeanthon, 2000; Mills *et al.*, 2004; Reysenbach & Cady, 2001). Paradoxically, despite their essential role in nutrient cycling, the planktonic communities inhabiting the vast biotope found between the photic region and the ocean floor are still poorly known. Only a few sparse molecular surveys of prokaryotic diversity in this biotope are available. A number of deep-sea-related archaeal and bacterial lineages have been identified in surveys carried out in 500-m- and 3000-m-deep Pacific and 1000-m-deep Atlantic samples (Fuhrman & Davis, 1997), and in 3000-m-deep Antarctic waters (López-García *et al.*, 2001a, b). The diversity found in those surveys was remarkably different from that observed in surface 16S rDNA libraries, in accordance with the idea that vertical stratification is a major determinant in the water column and with reports concerning more superficial waters (Giovannoni *et al.*, 1996; Gordon & Giovannoni, 1996).

Among the *Bacteria*, the *Deltaproteobacteria* appear to be one of the groups that increase in quantitative and qualitative representation in 16S rDNA libraries with depth. The *Deltaproteobacteria* groups organisms that can be classified into four major types according to their lifestyle (Seeger & Stetter, 1999). Some lineages are dissimilatory sulfate- or sulfur-reducers, being widespread in anoxic sediments all around the world. Many of these lineages, such as *Geobacter* species (Lovley *et al.*, 2004), are not only able to reduce sulfur species, but also oxidized metals such as Fe(III) or Mn(IV). Others are fermenters that live in syntrophic symbioses with hydrogen-consuming methanogens (e.g. *Syntrophus*, *Syntrophobacter*). Some others are unusual bacterial predators, such as *Bdellovibrio* spp. (Rendulic *et al.*, 2004). Finally, the myxobacteria constitute a unique group of heterotrophs displaying well developed cell-to-cell communication and complex developmental life cycles, which include cellular differentiation and multicellular stages (Dawid, 2000; Dworkin, 1996; Reichenbach, 1999). Although thought to be exclusively aerobic for a long time, anaerobic myxobacteria have been isolated from various soils, sediments and the subsurface that are able to respire aryl-compounds and, similarly to other *Deltaproteobacteria*, Fe(III) (He & Sanford, 2003; Sanford *et al.*, 2002). In the deep ocean, sulfate-reducing members of the *Deltaproteobacteria* are frequently detected in sediments, cold seeps and anoxic regions in the water column, such as the Cariaco basin (Inagaki *et al.*, 2002; Madrid *et al.*, 2001). *Bdellovibrio* and myxobacteria-related sequences are also well represented in deep-sea sediments around hydrothermal vents (López-García *et al.*, 2003). Among planktonic organisms, a novel lineage related to the *Deltaproteobacteria*, SAR324, was identified in lower-surface-level waters at the Sargasso Sea at a depth of 250 m (Wright *et al.*, 1997) that was also detected in far deeper waters (López-García *et al.*, 2001a).

We previously constructed a genomic library of plankton from a depth of 500 m in the size range 0.2–5 µm from the Antarctic Polar Front, from which two archaeal clones have been analysed previously (López-García *et al.*, 2004; Moreira *et al.*, 2004). We have extended this library and screened 3200 additional clones. The only two bacterial cosmid clones, designated DeepAnt-32C6 and DeepAnt-1F12, carrying 16S–23S rDNA intergenic spacers (ITSs) distinguishable from the host *Escherichia coli* strain belonged to the *Deltaproteobacteria*. We have sequenced and analysed the genome fragments borne by these clones and found them to be very divergent from known *Deltaproteobacteria*, defining a novel group branching as a sister group to the myxobacteria.

METHODS

Genomic library construction and screening. In this work, we enlarged an initial metagenomic library of 6107 cosmid clones of Antarctic plankton from a depth of 500 m that had been constructed using the Epicentre pWEB Cloning System (López-García *et al.*, 2004) with 3200 additional genomic clones (~150 Mbp environmental DNA). Clones were obtained using *E. coli* EPI305, which is genetically deficient in both recombination and restriction systems to avoid the rearrangement or loss of clones *in vivo* (Epicentre). The 3200 clones were PCR-screened for the presence of bacterial 16S rRNA genes. To differentiate the host *E. coli* 16S rDNA amplicons from those potentially present in its cosmids, we used the primers 1055F (5'-ATGGCTGTCGTCAGCTCGT-3') and 230R (5'-TGCCAAGGCATCCACCGT-3') allowing the amplification of the last third of the 16S rRNA gene plus the ITS separating it from the 23S rRNA gene. Two bacterial clones, designated DeepAnt-32C6 and DeepAnt-1F12, were detected among the 3200 clones in this way. 16S rDNAs were then amplified using B-27F (5'-AGAGTTT-GATCCTGGCTCAG-3') and 1492R (5'-GGTTACCTGTTCAG-CTT-3'), completely sequenced and compared by using BLAST (Altschul *et al.*, 1997) to database sequences.

Cosmid sequencing. DeepAnt-32C6 and DeepAnt-1F12 were fully sequenced (Genome Express, France). Cosmid DNA was purified from overnight 350 ml cultures. Shotgun libraries were then constructed by subcloning mechanically sheared DNA fragments (2.5–5 kbp) into pUC18. Plasmid inserts were sequenced using vector primers in 3730XL capillary sequencers (Applied Biosystems). After cleaning vector sequences, DeepAnt-32C6 yielded one contig of 36 661 bp with a mean coverage of 8.8×. In the case of DeepAnt-1F12, two contigs of 12 354 bp (9.5× mean coverage) and 26 067 bp (10.4× mean coverage) were initially obtained that were finally assembled into one single contig of 38 421 bp.

Gene annotation and analysis of sequence features. ORFs were characterized using the ORF-finder software at the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/gorf/gorf.html). Several criteria were used to define bona fide ORFs among all the potential ORFs detected. First, only sequences encoding peptides longer than 50 aa, preferentially non-overlapping, were retained. Second, when putative ORFs were detected in different reading frames, we selected those that had known homologues (as detected by BLAST searches) and third, in the case of various putative overlapping ORFs in different reading frames with no known homologues, we selected the ORFs that had the longest sequence. We used the programs PSI-BLAST, TBLASTN and BLASTP (Altschul *et al.*, 1997) to look for sequence similarity in each ORF product with published protein sequences. The analyses of predicted conserved domains and transmembrane segments were carried out using

CDART (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi; Geer *et al.*, 2002) and TMPRED (Hofmann & Stoffel, 1993), respectively. Predicted proteins were classified according to the Clusters of Orthologous Groups of proteins (COGs) database (www.ncbi.nlm.nih.gov/COG/; Tatusov *et al.*, 1997, 2000). The presence of tRNA genes was explored with the program tRNAscan (Lowe & Eddy, 1997). tRNAs were absent from both the DeepAnt-32C6 and DeepAnt-1F12 cosmids. The presence of tandem repeats was analysed with the Tandem Repeats Finder (<http://tandem.bu.edu/tools.html>; Benson, 1999) and Oligonucleotides Repeats Finder (<http://www.mgs.bionet.nsc.ru/mgs/programs/oligorep/InpForm.htm>). GC composition was studied with the GC Content Grapher (<http://plantst.sdsc.edu/plantst/html/geneGC.shtml>).

Phylogenetic analysis. In the case of 16S rRNA genes, the closest relatives to the DeepAnt-32C6 and DeepAnt-1F12 cosmids identified by BLASTN were incorporated into an alignment containing ~17 000 prokaryotic sequences. Sequences were aligned automatically using the program BABA (H. Philippe, personal communication) and the alignment was refined manually using the program ED of the MUST package (Philippe, 1993). Gaps and ambiguously aligned positions were excluded from phylogenetic tree reconstruction analyses. Preliminary trees with a large selection of species from the different bacterial divisions were constructed by neighbour-joining (using the MUST package) to identify the phylogenetic ascription of the two cosmids. More accurate trees including close relatives to our sequences were constructed by applying the Bayesian methods implemented in the program MRBAYES (Ronquist & Huelsenbeck, 2003) and maximum-likelihood using TREEFINDER (Jobb *et al.*, 2004). Maximum-likelihood phylogenetic reconstruction was done using a GTR+ Γ model of sequence evolution, using eight rate categories. Maximum-likelihood bootstrap analysis was carried out on 1000 replicates. In the case of proteins, the sequences homologous to the DeepAnt-32C6 and DeepAnt-1F12 ORFs identified by BLAST were retrieved from GenBank. Amino acid alignments for each ORF product were constructed using CLUSTALW (Thompson *et al.*, 1994) and inspected manually with the program ED of the MUST package (Philippe, 1993). All ambiguously aligned regions were excluded from phylogenetic analyses. For all alignments, preliminary neighbour-joining trees were constructed using MUST. Bayesian phylogenetic trees with a representative selection of species were then constructed using MRBAYES. All individual alignments and the corresponding phylogenetic trees are available upon request.

RESULTS AND DISCUSSION

Library extension, screening and identification of two divergent clones forming a novel lineage of the *Deltaproteobacteria*

For this work, we extended a metagenomic library of Antarctic mesopelagic pico- and small nanoplankton (López-García *et al.*, 2004; Moreira *et al.*, 2004) by generating 3200 additional clones. These were screened for the presence of bacterial 16S rRNA genes using specific primers allowing the amplification of the last third of the bacterial 16S rDNA plus the ITS (see Methods). Since the ITS is highly variable in sequence and length, clones yielding an amplification pattern different from that of *E. coli* alone would carry bacterial 16S rDNA-containing cosmids (Beja *et al.*, 2000b) although, obviously, the fraction of total 16S rRNA-containing clones in the library may be underestimated by this procedure. Only two bacterial clones,

DeepAnt-32C6 and DeepAnt-1F12, were identified using this strategy. Once purified, their full-length 16S rDNAs were amplified, sequenced and compared to sequences deposited in non-redundant and environmental databases by BLAST.

DeepAnt-32C6 and DeepAnt-1F12 appeared clearly related to deltaproteobacterial sequences, but sequence identity to their close relatives was very low. Only in the case of DeepAnt-1F12 did the closest 16S rDNA sequence in the databases exhibit more than 90 % identity. This closest match was the environmental sequence GR-WP33-58 (AJ296570), which was retrieved from a uranium waste pile (Selenska-Pobell, 2002). The second closest BLAST hit was only 89 % identical and corresponded to the highly versatile metal-reducer *Geobacter sulfurreducens* (Methe *et al.*, 2003). In the case of DeepAnt-32C6, the situation was even more extreme. The closest BLAST hit, *Pelobacter propionicus* (X70954), was only 89 % identical; the second closest hit corresponded again to *G. sulfurreducens* (88 % identity). No close relatives were observed among the environmental Sargasso Sea surface sequences released recently (Venter *et al.*, 2004); the closest hits to our DeepAnt clones identified in the Sargasso Sea database were proteobacterial sequences not belonging to the delta subclass (not shown). To determine the correct phylogenetic position of DeepAnt-32C6 and DeepAnt-1F12, we incorporated their 16S rDNA sequences and those from their closest BLAST-detected relatives to a prokaryotic alignment containing ~17 000 sequences, and we carried out various preliminary phylogenetic analyses using diverse widely distributed taxon samples. In all cases, the two DeepAnt clones grouped together with the environmental clone GR-WP33-58 within the *Deltaproteobacteria* and, although their position within this group was not clear, it tended to branch as a sister group to the myxobacteria (not shown). Therefore, we chose a representation of proteobacterial sequences including a broad sampling of myxobacteria and other *Deltaproteobacteria* to carry out more exhaustive phylogenetic analyses (Fig. 1). The degree of divergence of this novel clade is comparable to that existing between the different deltaproteobacterial orders. In addition to defined orders, we also included representative sequences of SAR324 (Wright *et al.*, 1997). Sequences belonging to this cluster had been identified at a depth of 3000 m in the water column of the Antarctic Polar Front, i.e. 2500 m deeper than the depth from which samples used to construct our genomic library were collected (López-García *et al.*, 2001a), suggesting that SAR324 was indeed specific to the aphotic water column. Furthermore, the group B sequences identified by Fuhrman & Davis (1997) in samples from depths of 1000 and 3000 m in the Pacific and Atlantic clearly belong to SAR324 (data not shown), but they were not included in this analysis because they are partial. Nevertheless, although SAR324 appears to be associated with mesopelagic and far deep-sea waters in various oceanic regions, the two DeepAnt clones identified in our genomic library were not related to it. On the contrary, they form a monophyletic cluster together

Fig. 1. Maximum-likelihood phylogenetic tree of 16S rRNA genes including a wide representation of *Deltaproteobacteria* to show the position of the novel myxobacteria-related group (MRG) revealed by the mesopelagic Antarctic clones DeepAnt-32C6 and DeepAnt-1F12. Several proteobacterial species were used as an outgroup. Bootstrap values higher than 50 % are shown at nodes.

with GR-WP33-58, branching as a sister group to the myxobacteria, although with only moderate support (Fig. 1).

Despite the possible underestimation of the diversity of bacterial 16S rDNA-containing clones in the library by the screening strategy used, the group defined by these two clones appears to be abundant in this particular location (Antarctic Polar Front) and depth (500 m), since DeepAnt-32C6 and DeepAnt-1F12 were the only two clones containing bacterial 16S rDNA genes among 3200 clones. The analysis of genomic libraries yields more approximate estimates of relative abundances than 16S rDNA libraries

because, even when inevitable biases can be introduced during DNA extraction and cloning, they escape PCR-derived bias. Despite the lack of extensive databases on deep-sea microbial diversity, the relatively high abundance of *Deltaproteobacteria* observed in our genomic library seems to be in agreement with existing data from deep-sea 16S rDNA libraries showing not only the presence of new deltaproteobacterial lineages (Fuhrman & Davis, 1997; López-García *et al.*, 2001a; Wright *et al.*, 1997), but also an increase in their relative proportion with depth. Thus, in a recent comparative study of 1438 bacterial 16S rDNAs from libraries obtained from surface (50 m depth) and deep

sea (2000 and 3000 m depth) in two oceanic locations, a significant and concordant increase of deltaproteobacterial clones was observed with depth. Deltaproteobacterial clones increased from 2.5 to 12.4% in the Ionian Sea at the Mediterranean basin, and from 1.4 to 12.3% in the North Atlantic close to Greenland (Zaballos *et al.*, 2006). Although these percentages should be viewed with caution, it is interesting to note that the proportion of deltaproteobacterial sequences obtained by an independent approach in the surface of the Sargasso Sea, around 2.5% (Venter *et al.*, 2004), is in fair agreement with the proportion of surface deltaproteobacterial clones in the other two locations.

From its intermediate phylogenetic position between versatile metal-reducers (such as *Geobacter*) and myxobacteria, inferring the general features and lifestyle of this novel myxobacteria-related group (MRG) is a highly speculative exercise. Nevertheless, although most known myxobacteria are heterotrophic aerobes, anaerobic Fe(III)-respiring myxobacteria have been described (Petrie *et al.*, 2003). It is therefore tempting to suggest that the ability to reduce not only sulfate, a property most likely ancestral to all *Deltaproteobacteria* (Woese, 1987), but also other oxidized metallic species, was ancestral to the myxobacteria and might also be found in the MRG. Clearly, the study of the full sequence of the two clones, DeepAnt-32C6 and DeepAnt-1F12, could provide information not only about genome organization and gene content of the MRG but, if genes involved in metabolic pathways were identified, also clues about their possible physiology. Hence, we completely sequenced the two genomic clones.

Genomic organization of DeepAnt-1F12 and DeepAnt-32C6

The sizes of the genomes of our two deltaproteobacterial clones, DeepAnt-1F12 and DeepAnt-32C6, were 38 421 and 36 589 bp, respectively. Both clones harboured a complete rRNA operon with 16S and 23S rRNA genes. However, tRNA genes, which are commonly found in the bacterial 16S–23S rDNA ITSs, including those of *Deltaproteobacteria* (Heidelberg *et al.*, 2004; Methe *et al.*, 2003; Rabus *et al.*, 2004; Rendulic *et al.*, 2004), were absent in both clones. In addition to the rRNA genes, we identified 15 ORFs in DeepAnt-1F12 and 18 in DeepAnt-32C6 (Fig. 2). None of these ORFs was shared by the two clones, but for most of them, clear homologues could be identified by BLAST searches in public databases (Tables 1 and 2). In fact, only three ORFs, one in DeepAnt-32C6 and two in DeepAnt-1F12, lacked close homologues in the databases. Gene density was much higher in DeepAnt-32C6 (with ITSs of 163 bp on average) than in DeepAnt-1F12 (with ITSs of 858 bp on average). Most ORFs in DeepAnt-1F12 (11 out of 15) and DeepAnt-32C6 (9 out of 16) were oriented in the same direction as the rRNA genes (Fig. 2).

Both clones differed significantly in their G + C content, with values of 59.06 mol% for DeepAnt-1F12 and

62.7 mol% for DeepAnt-32C6, although their rRNA genes showed very similar contents (54.67 and 54.90 mol%, respectively). The G + C graph for DeepAnt-1F12 showed a long region with a G + C content much higher than the mean, corresponding to ORFs 3 and 4, encoding two large (784 and 1293 aa) glycine-rich proteins (Fig. 2). On the contrary, the only region in DeepAnt-32C6 with a divergent, much lower G + C content corresponded to the rRNA operon. Interestingly, the G + C content of the protein-coding regions was very similar in both clones, 62.45 mol% in DeepAnt-1F12 and 63.93 mol% in DeepAnt-32C6. Therefore, the much lower mean G + C content of DeepAnt-1F12 is due to the relatively AT-rich intergenic regions, which are much larger in this clone, occupying approximately one-third of its total length.

Protein-coding genes

DeepAnt-1F12 and DeepAnt-32C6 carried a variety of genes encoding proteins with both informational and operational functions. Some of these proteins, particularly those of an informational nature (e.g. a DNA polymerase I subunit, ORF 8, in DeepAnt-32C6, or an RNA polymerase subunit, ORF 12, in DeepAnt-1F12), were highly conserved, and their phylogenetic analysis unambiguously confirmed the deltaproteobacterial affinity of the two clones (Tables 1 and 2, and data not shown). Other proteins were much less conserved and did not allow us to make robust phylogenetic inferences. Finally, a few proteins were conserved and gave a good phylogenetic signal, but their position in phylogenetic trees was very distant from other *Deltaproteobacteria*. This could be explained by hidden paralogies (i.e. independent gene duplications and losses), but it would require the occurrence of an improbably huge number of paralogues (Koonin, 2003). Therefore, we interpreted these phylogenetic positions as possible evidence for horizontal gene transfer (HGT) events. In most cases, these putative HGTs involved both low-G + C and high-G + C Gram-positives (Tables 1 and 2). One clear example is that of DeepAnt-32C6 ORF 3 encoding an adenosylmethionine-8-amino-7-oxononanoate aminotransferase that is closely related to high-G + C Gram-positive homologues with a strong statistical support (supplementary Fig. S1, available with the online version of this paper). In other, less frequent cases, transferred genes appear to have been acquired from other bacterial groups (Tables 1 and 2). An interesting example corresponds to DeepAnt-1F12 ORF 8, encoding a conserved transposase domain. Transposases are enzymes involved in the mobilization of DNA segments, being important agents of gene transfer (Osborn & Boltner, 2002). The transposase present in our clone is widespread in *Gammaproteobacteria*, but it is also found in one alphaproteobacterium (*Sphingobium herbicidovorans*) and one betaproteobacterium (*Delftia acidovorans*), and in one high-G + C Gram-positive (*Arthrobacter aurescens*) species with an almost identical sequence, suggesting a very recent acquisition from the *Gammaproteobacteria* (supplementary Fig. S2, available with the online version of this paper).

Fig. 2. Schematic gene organization of DeepAnt-1F12 and DeepAnt-32C6 genome fragments. Dark grey bars, 16S rDNA; light grey bars, 23S rDNA; solid arrows, ORFs with homologues in databases; dashed arrows, hypothetical ORFs. The graphs displayed beneath the gene maps indicate the variation in G + C content along the two genomic clones.

However, in the case of DeepAnt-1F12, the gene is truncated and very divergent, indicating that this was a more ancient acquisition and that it has probably become a pseudogene.

Interestingly, DeepAnt-1F12 has one gene (ORF 10) encoding a Cys-rich protein that was clearly related to eukaryotic homologues. Although default BLAST searches against the non-redundant database in GenBank gave exclusively eukaryotic hits, in-depth searches in bacterial incomplete genome sequences allowed the identification of a homologue in the myxobacterium *Myxococcus xanthus* (Fig. 3). This might be indicative of an actual closer relationship between the new deltaproteobacterial group defined by our DeepAnt clones, GR-WP33-58 and the myxobacteria, although DeepAnt-1F12 was not a sister to *M. xanthus* in the phylogenetic analysis. In fact, the gene is fast-evolving and the statistical support for the separation of the two sequences was very low. In addition, this gene is present in multiple paralogous copies in eukaryotes (e.g. at least six paralogues in *Homo sapiens*; see Fig. 3), so that it is possible that the sequences found in DeepAnt-1F12 and *M. xanthus* belong to different paralogous families. In addition, BLAST searches in the Sargasso Sea environmental database identified two closely homologous sequences, IBEA_CTG_UEAYZ23TF and IBEA_CTG_UAANF02TF, that matched with only eukaryotic proteins (data not shown). However, both of these Sargasso Sea sequences are

truncated and, since these clones correspond to small DNA fragments (<1000 bp) that do not carry other phylogenetic markers, it is very difficult to know the actual organismal origin. Nevertheless, IBEA_CTG_UAANF02TF was so similar to sea urchin sequences that an echinodermal source is highly probable. The surprising similarity of DeepAnt-1F12 ORF 10 with several eukaryotic proteins reinforces the idea that this MRG group is truly related to the myxobacteria, since the latter share a number of features, including many genes, with eukaryotes, mostly related to cell-cell communication and differentiation (Dworkin, 1996). This gene would therefore belong to the pool of eukaryotic-like genes carried by myxobacteria. The closest BLAST homologues to DeepAnt-1F12 ORF 10, i.e. EtMIC4 and EmTPF250 from the apicomplexan parasites *Eimeria tenella* and *Eimeria maxima* (Tomley *et al.*, 2001; Witcombe *et al.*, 2003), have an unknown function. Nevertheless, all the homologues found contain several repetitions of a Ca^{2+} -binding epidermal growth factor (EGF) domain. EGFs are involved in protein-protein interaction and are found in numerous eukaryotic extracellular and membrane proteins important for motility, signalling, adhesion and development (Beckingham *et al.*, 1998). Hence, the protein in myxobacteria and in this novel deltaproteobacterial group could be a membrane protein (ORF 10 has two predicted strong transmembrane domains) with Ca^{2+} -binding activity. Based on this, it would be tempting to speculate that

Table 1. Predicted rRNA- and protein-coding genes in the cosmid DeepAnt-1F12

ORF	Position	Frame	Protein size (aa)	Predicted function	BLAST best hit (E-value)	Phylogenetic affinity*	Comments
1	<1–523†	+2	>174†	FAD-dependent dehydrogenase (COG2509)	<i>Cytophaga hutchinsonii</i> (3e-34)	Bacteria	
16S rDNA	821–2353	+		16S rRNA	Uncultured bacterium GR-WP33-58 (92 % identity)	Deltaproteobacteria	
23S rDNA	2955–5995	+		23S rRNA	<i>Geobacter sulfurreducens</i> (89 % identity)	Deltaproteobacteria	
2	6922–9762	+1	946	Conserved protein of unknown function DUF229 (COG1368)	<i>Rubrobacter xylophilus</i> (7e-14)	Bacteria	Members of this family are uncharacterized 500–1200 aa proteins that share a long conserved region. Some are similar to sulfatases or to phosphoglycerol transferase and related proteins, alkaline phosphatase superfamily. Seven strong transmembrane helices.
3	10 072–12 426	+1	784	Gly-X-X-Gly-rich protein	<i>Trichodesmium erythraeum</i> (1e-176)	Bacteria	
4	11 957–15 838	+2	1293	Gly-X-X-Gly-rich protein	<i>Trichodesmium erythraeum</i> (0-0)	Bacteria	
5	19 423–20 283	+1	286	Esterase/lipase (COG0657)	<i>Clostridium perfringens</i> (6e-14)	Proteobacteria	Two strong transmembrane helices
6	20 412–21 479	+3	356	Hypothetical	<i>Tetraodon nigroviridis</i> (0-072)	Unknown	
7	22 984–25 134	+1	716	Hypothetical	<i>Tetraodon nigroviridis</i> (0-001)	Unknown	
8	25 505–25 765	+2	86	Transposase, partial sequence	<i>Pseudomonas putida</i> (1e-09)	Gammaproteobacteria	Probable HGT
9	26 214–26 849	+3	211	Acetyltransferase, GNAT family (COG3393)	<i>Bacillus cereus</i> (3e-25)	Low-G+C Gram-positives	Probable HGT
10	27 413–30 604	+2	1063	Similar to eukaryotic structural proteins	<i>Eimeria tenella</i> (1e-118)	Eukaryotes	Homologues are found only in eukaryotes. Contains seven Ca-binding Epidermal Growth Factor (EGF) domains
11	31 353–31 748	–2	131	Transmembrane domain 2 (TM2) protein	<i>Leptospira interrogans</i> (5e-15)	Spirochaetes	Two strong transmembrane helices. Probable HGT.
12	32 089–35 559	–2	1156	Multidomain protein	<i>Bdellovibrio bacteriovorus</i> HD100 (1e-123)	Deltaproteobacteria	Contains three conserved domains: VIT (vault protein inter-alpha-trypsin domain), vWFA (von Willebrand factor type A domain) and RpoN (COG1508, DNA-directed RNA polymerase specialized sigma subunit)

Table 1. cont.

ORF	Position	Frame	Protein size (aa)	Predicted function	BLAST best hit (E-value)	Phylogenetic affinity*	Comments
13	35 698–36 204	+1	168	RpoE, DNA-directed RNA polymerase (COG1595)	<i>Bradyrhizobium japonicum</i> (1e-04)	Proteobacteria	
14	36 674–37 066	–3	130	Similar to conserved bacterial hypothetical proteins	<i>Xanthomonas campestris</i> (0.004)	Bacteria	
15	37 480–37 833	–1	117	Similar to conserved bacterial hypothetical proteins	<i>Mycobacterium avium</i> (1e-12)	High-G + C Gram-positives	Probable HGT

*Inferred by phylogenetic tree reconstruction.

†This ORF is partial at its 3' end.

this novel deltaproteobacterial group shares some of the complex features of social behaviour and development typical of the myxobacteria.

ORF 12 of DeepAnt-1F12 encoded another unusual protein. It was a multidomain polypeptide containing a VIT (vault protein inter-alpha-trypsin domain), a vWFA (von Willebrand factor type A domain) and an RpoN (DNA-directed RNA polymerase specialized sigma subunit). The RpoN region was highly conserved and phylogenetic analysis corroborated its deltaproteobacterial affinity (not shown). RpoN is an essential protein for RNA transcription universally distributed in bacteria. VIT domains are involved in the formation of 'vaults', which are highly conserved ribonucleoprotein particles with an undefined function, frequently found in eukaryotes (Kickhoefer *et al.*, 2002). VIT domains are frequently associated with vWFA domains, which are transmembrane peptides. Although previously thought to be exclusive to eukaryotic proteins, the association of VIT-vWFA domains could also be found by BLAST searches in many bacterial proteins, but the combination of these domains with RpoN appears to be unique.

Many of the protein-coding genes found in our two metagenomic clones encoded proteins involved in house-keeping functions universally distributed in prokaryotes (e.g. DNA replication, RNA synthesis), so that they were not informative about the possible physiological characteristics of the new MRG defined by DeepAnt-1F12 and DeepAnt-32C6. One interesting exception was ORF 14 of DeepAnt-32C6, which encoded a thiamine-dependent acetolactate synthase. This enzyme is found in several strictly or facultatively anaerobic bacteria, including *Deltaproteobacteria* (Lovley *et al.*, 1995) and *Archaea*, and is required for 2,3-butanediol fermentation, one of the known pathways for glucose fermentation (Syu, 2001). 2,3-Butanediol, CO₂, H₂, ethanol and lactate are the major products of this fermentative pathway, which strongly decreases acid formation (butanediol is neutral) in comparison with other glucose fermentation pathways that typically produce much larger amounts of acid (such as acetate). The presence of this gene in DeepAnt-32C6 suggests that this deltaproteobacterium is able to carry out 2,3-butanediol fermentation and, therefore, anaerobic energy metabolism. The presence of fermentative micro-organisms in the deep-sea might be surprising, since this region of the ocean is characterized by a large amount of dissolved O₂. However, the recent analysis of the complete genome sequence of the deep-sea gammaproteobacterium *Idiomarina loihiensis* has revealed that this species relies primarily on amino acid fermentation as a source of both carbon and energy (Hou *et al.*, 2004), suggesting that fermentation may be important even in oxygenated deep-sea waters. Interestingly, ORF 13, immediately upstream and with a 127 bp overlap with the thiamine-dependent acetolactate synthase gene (Table 2), encoded a multidomain protein that is most probably involved in signal transduction. It contained a conserved

Table 2. Predicted RNA- and protein-coding genes in the cosmid DeepAnt-32C6

ORF	Position	Frame	Protein size (aa)	Predicted function	BLAST best hit (E-value)	Phylogenetic affinity*	Comments
16S rDNA	913–2447	+		16S rRNA	<i>Pelobacter propionicus</i> (88 % identity)	<i>Deltaproteobacteria</i>	
23S rDNA	2955–5995	+		23S rRNA	<i>Geobacter sulfurreducens</i> (89 % identity)	<i>Deltaproteobacteria</i>	
1	6531–7784	–3	417	Esterase (COG1680)	<i>Acinetobacter</i> sp. ADP1 (1e-71)	High-G + C Gram-positives	Two strong transmembrane helices
2	8551–10 062	+1	503	NADP-dependent aldehyde dehydrogenase	<i>Bdellovibrio bacteriovorus</i> (1e-133)	<i>Deltaproteobacteria</i>	Three strong transmembrane helices
3	10 119–11 438	+3	439	BioA, adenosylmethionine-8-amino-7-oxononanoate aminotransferase (COG0161)	<i>Streptomyces coelicolor</i> (8e-91)	High-G + C Gram-positives	
4	11 537–12 427	–1	296	Amidohydrolase (COG0388)	<i>Brevibacillus agri</i> (1e-99)	High-G + C Gram-positives	
5	12 549–13 571	+3	340	Dehydrogenases with pyrrolo-quinoline quinone (PQQ) as cofactor	<i>Chloroflexus aurantiacus</i> (6e-21)	<i>Bacteria</i>	Three strong transmembrane helices
6	13 646–14 647	+2	333	Tetratricopeptide repeat (TPR) domain protein	<i>Nostoc punctiforme</i> (2e-19)	<i>Deltaproteobacteria</i>	Contains six TPR repeats
7	14 736–17 615	+3	959	Uncharacterized protein conserved in bacteria (COG4676)	' <i>Microbulbifer degradans</i> ' (3e-05)	<i>Proteobacteria</i>	Four strong transmembrane helices
8	17 839–20 412	+1	857	DNA polymerase I (PolI)	<i>Azotobacter vinelandii</i> (1e-165)	<i>Deltaproteobacteria</i>	
9	20 475–21 068	+3	197	RNA pseudouridylate synthase (COG0564)	<i>Geobacter sulfurreducens</i> (5e-16)	<i>Deltaproteobacteria</i>	
10	21 116–22 693	–1	525	Spermidine synthase (COG4262)	' <i>Dechloromonas aromatica</i> ' (8e-85)	<i>Bacteria</i>	Seven strong transmembrane helices
11	22 694–23 593	–1	299	Hypothetical transmembrane protein	<i>Rhodopirellula baltica</i> (2e-11)	<i>Bacteria</i>	
12	23 674–24 276	–2	200	Hypothetical transmembrane protein	' <i>Dechloromonas aromatica</i> ' (1e-14)	<i>Bacteria</i>	Two strong transmembrane helices
13	24 328–27 195	–2	955	Multidomain signal transduction protein (COG5001)	<i>Rubrobacter xylanophilus</i> (1e-151)	High-G + C Gram-positives	Contains three domains: PAS (sensor for oxygen in signal transduction), DUF1 (domain of unknown function 1, likely to participate in prokaryotic signalling processes) and DUF2 (domain of unknown function 2, occurring exclusively in bacteria)

Table 2. cont.

ORF	Position	Frame	Protein size (aa)	Predicted function	BLAST best hit (E-value)	Phylogenetic affinity*	Comments
14	27068–29110	+2	680	IlvB, acetolactate synthase requiring thiamine pyrophosphate (COG0028)	<i>Methanococcus jannaschii</i> (2e-77)	High-G+C Gram-positives	Three strong transmembrane helices
15	29233–31710	+1	825	Hypothetical	<i>Photobacterium profundum</i> (0-011)	Bacteria	Five strong transmembrane helices
16	31771–33171	-2	466	Pyruvate kinase	<i>Thermus thermophilus</i> (1e-113)	Cyanobacteria	Two strong transmembrane helices
17	33392–33937	+2	181	Hypothetical protein	<i>Rhodospirillum rubrum</i> (0-007)	Bacteria	
18	33925–34203	+1	92	Hypothetical protein	<i>Myxococcus xanthus</i> (0-19)	Bacteria	

*Inferred by phylogenetic tree reconstruction.

PAS domain, which is known to bind a variety of ligands and to act as a sensor for oxygen (Makino *et al.*, 2001; Zhulin *et al.*, 1997). The presence of this conserved signal transduction protein in DeepAnt-32C6 suggests that this deltaproteobacterium is indeed able to sense changes in environmental O₂ concentration and regulate gene expression accordingly. Interestingly, both this protein and the adjacent thiamine-dependent acetolactate synthase appeared to have been acquired by HGT from high-G + C Gram-positive bacteria, as deduced from phylogenetic analyses (not shown).

Another gene in DeepAnt-32C6 with potential metabolic information is ORF 1, encoding an esterase (Table 2). Closely related homologues were found in a variety of bacterial species. They are involved in the extracellular degradation of complex compounds containing ester bonds and cyclic structures to produce simpler molecules, which are subsequently incorporated and metabolized in the cell (e.g. cyclododecanone; Kostichka *et al.*, 2001). This might also be the case for DeepAnt-32C6 ORF 1, since it contains two predicted strong transmembrane helices, suggesting that it is most probably membrane-bound. Therefore, the presence of this gene, as well as ORFs 13 and 14, suggested that this species is capable of both anaerobic growth and degradation of complex extracellular compounds. Since, as mentioned above, deep-sea water is rather oxygenated and poor in dissolved organic matter, one possibility could be that this deltaproteobacterium thrives, at least temporarily, within the suspended particles that form particulate organic matter, where anaerobic microniches rich in organic matter are frequent (del Giorgio & Duarte, 2002).

Conclusions

Our metagenomic analysis of a mesopelagic library has revealed the existence of a novel deltaproteobacterial group, the MRG, in marine plankton. The number of MRG deltaproteobacterial clones found in our library suggests that this group may be relatively abundant in deep-sea plankton. The phylogenetic position of the MRG is very interesting; since they branch at the base of the myxobacteria (Fig. 1), they could define an intermediate group between the very complex myxobacteria and the classical sulfate-reducing *Deltaproteobacteria*. Although the analysis of a small number of genes is insufficient for the inference of metabolic traits, our data suggest that at least one of the two deep-sea MRG *Deltaproteobacteria* might be capable of anaerobic growth based on fermentation. This emphasizes the necessity of further studies of the deep-sea microbial community, which appears to have a taxonomic and quantitative composition very different from that of the surface community and that may display metabolic properties (e.g. carbon and energy metabolism based on fermentation) relatively unexpected from the overall physico-chemical features of their biotope (Hou *et al.*, 2004).

Fig. 3. Bayesian phylogenetic tree showing the position of the Cys-rich predicted DeepAnt-1F12 ORF 10 and its eukaryotic homologues. Posterior probabilities are indicated at nodes.

ACKNOWLEDGEMENTS

We are grateful to M. Zaballos for sharing unpublished information with us. This work was financed by the European Commission Project GEMINI (QLK3-CT-2002-02056), the EC Network of Excellence *Marine Genomics Europe*, and the French CNRS – Ministère de la Recherche program *Séquençage à Grande Echelle*.

REFERENCES

- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Beckingham, K., Lu, A. Q. & Andruss, B. F. (1998). Calcium-binding proteins and development. *Biometals* **11**, 359–373.
- Beja, O., Aravind, L., Koonin, E. V. & 9 other authors (2000a). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906.
- Beja, O., Suzuki, M. T., Koonin, E. V. & 9 other authors (2000b). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**, 516–529.
- Beja, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. (2001). Proteorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789.
- Benson, G. (1999). Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res* **27**, 573–578.
- Dawid, W. (2000). Biology and global distribution of myxobacteria in soils. *FEMS Microbiol Rev* **24**, 403–427.
- De La Torre, J. R., Christianson, L. M., Beja, O., Suzuki, M. T., Karl, D. M., Heidelberg, J. & DeLong, E. F. (2003). Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci U S A* **100**, 12830–12835.
- del Giorgio, P. A. & Duarte, C. M. (2002). Respiration in the open ocean. *Nature* **420**, 379–384.
- DeLong, E. F. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci U S A* **89**, 5685–5689.
- DeLong, E. F. (2001). Microbial seascapes revisited. *Curr Opin Microbiol* **4**, 290–295.
- Dworkin, M. (1996). Recent advances in the social and developmental biology of the myxobacteria. *Microbiol Rev* **60**, 70–102.
- Fuhrman, J. A. & Davis, A. A. (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, J. A., McCallum, K. & Davis, A. A. (1992). Novel major archaeobacterial group from marine plankton. *Nature* **356**, 148–149.
- Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. (2002). CDART: protein homology by domain architecture. *Genome Res* **12**, 1619–1623.
- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63.
- Giovannoni, S. J., Rappe, M. S., Vergin, K. L. & Adair, N. L. (1996). 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the green non-sulfur bacteria. *Proc Natl Acad Sci U S A* **93**, 7979–7984.
- Gordon, D. A. & Giovannoni, S. J. (1996). Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific oceans. *Appl Environ Microbiol* **62**, 1171–1177.
- Hallam, S. J., Putnam, N., Preston, C. M., Detter, J. C., Rokhsar, D., Richardson, P. M. & DeLong, E. F. (2004). Reverse methanogenesis:

testing the hypothesis with environmental genomics. *Science* **305**, 1457–1462.

He, Q. & Sanford, R. A. (2003). Characterization of Fe(III) reduction by chlororespiring *Anaeromyxobacter dehalogenans*. *Appl Environ Microbiol* **69**, 2712–2718.

Heidelberg, J. F., Seshadri, R., Haveman, S. A. & 32 other authors (2004). The genome sequence of the anaerobic, sulfate-reducing bacterium *Desulfovibrio vulgaris* Hildenborough. *Nat Biotechnol* **22**, 554–559.

Hofmann, K. & Stoffel, W. (1993). TMbase – a database of membrane-spanning protein segments. *Biol Chem* **347**, 166.

Hou, S., Saw, J. H., Lee, K. S. & 19 other authors (2004). Genome sequence of the deep-sea gamma-proteobacterium *Idiomarina loihiensis* reveals amino acid fermentation as a source of carbon and energy. *Proc Natl Acad Sci U S A* **101**, 18036–18041.

Inagaki, F., Sakihama, Y., Inoue, A., Kato, C. & Horikoshi, K. (2002). Molecular phylogenetic analyses of reverse-transcribed bacterial rRNA obtained from deep-sea cold seep sediments. *Environ Microbiol* **4**, 277–286.

Jeanthon, C. (2000). Molecular ecology of hydrothermal vent microbial communities. *Antonie Van Leeuwenhoek* **77**, 117–133.

Jobb, G., von Haeseler, A. & Strimmer, K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**, 18.

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

Kickhoefer, V. A., Poderycki, M. J., Chan, E. K. & Rome, L. H. (2002). The La RNA-binding protein interacts with the vault RNA and is a vault-associated protein. *J Biol Chem* **277**, 41282–41286.

Koonin, E. V. (2003). Horizontal gene transfer: the path to maturity. *Mol Microbiol* **50**, 725–727.

Kostichka, K., Thomas, S. M., Gibson, K. J., Nagarajan, V. & Cheng, Q. (2001). Cloning and characterization of a gene cluster for cyclododecanone oxidation in *Rhodococcus ruber* SC1. *J Bacteriol* **183**, 6478–6486.

López-García, P., López-López, A., Moreira, D. & Rodríguez-Valera, F. (2001a). Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *FEMS Microbiol Ecol* **36**, 193–202.

López-García, P., Moreira, D., López-López, A. & Rodríguez-Valera, F. (2001b). A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environ Microbiol* **3**, 72–78.

López-García, P., Duperron, S., Philippot, P., Foriel, J., Susini, J. & Moreira, D. (2003). Bacterial diversity in hydrothermal sediment and epsilonproteobacterial dominance in experimental microcolonizers at the Mid-Atlantic Ridge. *Environ Microbiol* **5**, 961–976.

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.

Lovley, D. R., Phillips, E. J., Lonergan, D. J. & Widman, P. K. (1995). Fe(III) and S₀ reduction by *Pelobacter carbinolicus*. *Appl Environ Microbiol* **61**, 2132–2138.

Lovley, D. R., Holmes, D. E. & Nevin, K. P. (2004). Dissimilatory Fe(III) and Mn(IV) reduction. *Adv Microb Physiol* **49**, 219–286.

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

Madrid, V. M., Taylor, G. T., Scranton, M. I. & Chistoserdov, A. Y. (2001). Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin. *Appl Environ Microbiol* **67**, 1663–1674.

Makino, Y., Cao, R., Svensson, K., Bertilsson, G., Asman, M., Tanaka, H., Cao, Y., Berkenstam, A. & Poellinger, L. (2001). Inhibitory PAS domain protein is a negative regulator of hypoxia-inducible gene expression. *Nature* **414**, 550–554.

Methe, B. A., Nelson, K. E., Eisen, J. A. & 31 other authors (2003). Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science* **302**, 1967–1969.

Mills, H. J., Martinez, R. J., Story, S. & Sobecky, P. A. (2004). Identification of members of the metabolically active microbial populations associated with *Beggiatoa* species mat communities from gulf of Mexico cold-seep sediments. *Appl Environ Microbiol* **70**, 5447–5458.

Moreira, D., Rodríguez-Valera, F. & Lopez-Garcia, 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.

Morris, R. M., Rappe, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A. & Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810.

Osborn, A. M. & Boltner, D. (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**, 202–212.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740.

Petrie, L., North, N. N., Dollhopf, S. L., Balkwill, D. L. & Kostka, J. E. (2003). Enumeration and characterization of iron(III)-reducing microbial communities from acidic subsurface sediments contaminated with uranium(VI). *Appl Environ Microbiol* **69**, 7467–7479.

Philippe, H. (1993). MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res* **21**, 5264–5272.

Rabus, R., Ruepp, A., Frickey, T. & 15 other authors (2004). The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ Microbiol* **6**, 887–902.

Rappé, M. S. & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**, 369–394.

Reichenbach, H. (1999). The ecology of the myxobacteria. *Environ Microbiol* **1**, 15–21.

Rendulic, S., Jagtap, P., Rosinus, A. & 10 other authors (2004). A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science* **303**, 689–692.

Reysenbach, A. L. & Cady, S. L. (2001). Microbiology of ancient and modern hydrothermal systems. *Trends Microbiol* **9**, 79–86.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Sanford, R. A., Cole, J. R. & Tiedje, J. M. (2002). Characterization and description of *Anaeromyxobacter dehalogenans* gen. nov., sp. nov., an aryl-halorespiring facultative anaerobic myxobacterium. *Appl Environ Microbiol* **68**, 893–900.

Segerer, A. H. & Stetter, K. O. (1999). The order *Sulfolobales*. In *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*, <http://141.150.157.117:8080/prokPUB/chaprender/jsp/showchap.jsp?chapnum=29>. Edited by M. Dworkin. New York: Springer.

Selenska-Pobell, S. (2002). Diversity and activity of bacteria in uranium mining waste piles. In *Interactions of Microorganisms with Radionuclides*, pp. 225–254. Edited by M. J. Keith-Roach & F. R. Livens. Amsterdam: Elsevier.

Syu, M. J. (2001). Biological production of 2,3-butanediol. *Appl Microbiol Biotechnol* **55**, 10–18.

- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631–637.
- Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.
- Tomley, F. M., Billington, K. J., Bumstead, J. M., Clark, J. D. & Monaghan, P. (2001). EtMIC4: a microneme protein from *Eimeria tenella* that contains tandem arrays of epidermal growth factor-like repeats and thrombospondin type-I repeats. *Int J Parasitol* **31**, 1303–1310.
- Tyson, G. W., Chapman, J., Hugenholtz, P. & 7 other authors (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43.
- Venter, J. C., Remington, K., Heidelberg, J. F. & 20 other authors (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74.
- Witcombe, D. M., Belli, S. I., Wallach, M. G. & Smith, N. C. (2003). Molecular characterisation of EmTFP250: a novel member of the TRAP protein family in *Eimeria maxima*. *Int J Parasitol* **33**, 691–702.
- Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51**, 221–271.
- Wright, T. D., Vergin, K. L., Boyd, P. W. & Giovannoni, S. J. (1997). A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Appl Environ Microbiol* **63**, 1441–1448.
- Zaballos, M., López-López, A., Ovreas, L., Galán Bartual, S., D'Auria, G., Legault, B., Alba, J. C., Pushker, R., Daaeand, F. L. & Rodríguez-Valera, F. (2006). Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. *FEMS Microb Ecol* (in press).
- Zhulin, I. B., Taylor, B. L. & Dixon, R. (1997). PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. *Trends Biochem Sci* **22**, 331–333.