The ISME Journal (2011) 5, 285–304

www.nature.com/ismej

ORIGINAL ARTICLE

Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara

Achim Quaiser[1,2], Yvan Zivanovic[3], David Moreira[1] and Purificación López-García[1]
[1]*Unité d'Ecologie, Systématique et Evolution, CNRS UMR8079, Université Paris-Sud 11, Orsay, France;* [2]*Ecosystèmes, Biodiversité, Evolution, CNRS UMR 6553, Université de Rennes 1, Rennes, France and* [3]*Institut de Génétique et Microbiologie, UMR C8621 Université Paris-Sud 11, Orsay, France*

To extend comparative metagenomic analyses of the deep-sea, we produced metagenomic data by direct 454 pyrosequencing from bathypelagic plankton (1000 m depth) and bottom sediment of the Sea of Marmara, the gateway between the Eastern Mediterranean and the Black Seas. Data from small subunit ribosomal RNA (SSU rRNA) gene libraries and direct pyrosequencing of the same samples indicated that Gamma- and Alpha-proteobacteria, followed by Bacteroidetes, dominated the bacterial fraction in Marmara deep-sea plankton, whereas Planctomycetes, Delta- and Gamma-proteobacteria were the most abundant groups in high bacterial-diversity sediment. Group I Crenarchaeota/Thaumarchaeota dominated the archaeal plankton fraction, although group II and III Euryarchaeota were also present. Eukaryotes were highly diverse in SSU rRNA gene libraries, with group I (Duboscquellida) and II (Syndiniales) alveolates and Radiozoa dominating plankton, and Opisthokonta and Alveolates, sediment. However, eukaryotic sequences were scarce in pyrosequence data. Archaeal *amo* genes were abundant in plankton, suggesting that Marmara planktonic Thaumarchaeota are ammonia oxidizers. Genes involved in sulfate reduction, carbon monoxide oxidation, anammox and sulfatases were over-represented in sediment. Genome recruitment analyses showed that *Alteromonas macleodii* 'surface ecotype', *Pelagibacter ubique* and *Nitrosopumilus maritimus* were highly represented in 1000 m-deep plankton. A comparative analysis of Marmara metagenomes with ALOHA deep-sea and surface plankton, whale carcasses, Peru subsurface sediment and soil metagenomes clustered deep-sea Marmara plankton with deep-ALOHA plankton and whale carcasses, likely because of the suboxic conditions in the deep Marmara water column. The Marmara sediment clustered with the soil metagenome, highlighting the common ecological role of both types of microbial communities in the degradation of organic matter and the completion of biogeochemical cycles.
The ISME Journal (2011) **5,** 285–304; doi:10.1038/ismej.2010.113; published online 29 July 2010
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: deep-sea; anaerobic respiration; carbon fixation; carbon cycle; sulfate reduction; ammonia oxidation

Introduction

Marine waters and their underlying sediments constitute the largest portion of the biosphere. However, despite they are key for biogeochemical cycling in our planet, the composition of marine microbial communities, their metabolic potential and activities and their interactions with the environment remain poorly understood. A considerable effort has been devoted though in past years to study planktonic communities inhabiting euphotic layers by molecular and genomics-based methods that complement classical cultivation-based approaches. In fact, surface seawaters have been an environment of choice for pioneer studies of microbial diversity based on the amplification, cloning and sequencing of small subunit ribosomal RNA (SSU rRNA) genes (Schmidt *et al.*, 1991), as well as for more modern metagenomic (Venter *et al.*, 2004; DeLong *et al.*, 2006; Rusch *et al.*, 2007; Feingersch *et al.*, 2009) and metatranscriptomic (Frias-Lopez *et al.*, 2008; Poretsky *et al.*, 2009; Shi *et al.*, 2009) analyses. Knowledge about deep-sea planktonic communities is much more fragmentary, and the situation is even worse for deep-sea sediments, as with the exception of cold seeps, hydrothermal sediments or whale falls, which have raised some

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 4 March 2010; revised 18 May 2010; accepted 13 June 2010; published online 29 July 2010

attention because of their increased diversity and productivity (Jorgensen and Boetius, 2007), typical sea-bottom sediments have been barely explored. A metagenomic study of subsurface sediments at the Peru margin from 1 to 50 m below the seafloor (bottom depth 150.5 m; Biddle, personal communication) was only recently published (Biddle *et al.*, 2008). So far, metagenomic analyses targeting globally deep-sea planktonic communities have been carried out in two locations. DeLong *et al.* (2006) studied picoplankton at different depths from surface to 4000 m at the ALOHA station in the North Pacific subtropical gyre by constructing fosmid libraries and sequencing the insert extremities of randomly chosen clones, which yielded an average of 8–10 Mbp of sequence per library. A similar approach was used to study plankton coming from 3000 m depth from the Ionian Sea in the Mediterranean basin, which also produced ~10 Mbp of sequence (Martin-Cuadrado *et al.*, 2007). More recently, ~200 Mbp of sequence were obtained from a shotgun library for the 4000 m depth at the ALOHA station (Konstantinidis *et al.*, 2009). Comparing metagenomic sequences from different depths in the water column is helpful to determine the genes and the potential metabolic functions associated to stratified marine communities. Conversely, comparing metagenomic data from similar depth and different geographic locations can inform about the influence of local parameters and/or geographic distance on the composition and metabolic capabilities of microbial communities. Deep-sea (below 1000 m) planktonic environments are generally considered to be relatively uniform and stable in terms of microbial diversity, as microbial dispersal is high and conditions are overall relatively similar (oligotrophy, high pressure and low temperature). However, different water masses are endowed with different physico-chemical characteristics. Such differences are particularly important in close seas, much more influenced by coastal input and local features. For example, Mediterranean waters are very different from open oceanic waters, and this seems to be reflected at the metagenome level (Martin-Cuadrado *et al.*, 2007).

The Sea of Marmara has a maximum depth of ~1300 m and is the gateway between the Mediterranean Sea and the Black Sea, to which it is connected through the straits of the Dardanelles (65 m sill depth) and the Bosphorus (35 m sill depth), respectively. The Mediterranean and the Black seas have very different average salinities (38‰ and 18‰, respectively), which together with the shallow sill depths of the two straits, results in a stably stratified Marmara water column with a defined halocline at ~25 m of depth. The strong stratification limits vertical heat exchange and ventilation of subhalocline waters, leading to an almost constant temperature of 14.5 °C below the halocline and to very low concentrations of dissolved oxygen (30–50 $\mu mol\,kg^{-1}$) compared with surface waters, almost ten times as oxygenated (Ünlüata *et al.*, 1990; Beşiktepe *et al.*, 1994). The Sea of Marmara is subjected to a two-layer flow regime. On one hand, it is fed by the surface influx of brackish, nutrient-rich waters from the Black Sea, whereas there is an outflow of salty waters through the Bosphorus undercurrent. Biologically available nutrients imported from the Black Sea are rapidly consumed by photosynthetic organisms on surface waters, leading to a net export of particulate nutrients to deeper layers. On the other hand, Marmara receives the influx of salty (nearly 39‰), highly oxygenated, but nitrate- and phosphate-depleted Mediterranean (Aegean) waters. These salty, denser waters sink toward the bottom of the Marmara basin, contributing to the strong stratification (Ünlüata *et al.*, 1990; Beşiktepe *et al.*, 1994).

During the past 160 000 years, Marmara was disconnected from the Mediterranean several times when the eustatic sea level dropped below its outlet and evolved into a lake with freshened waters, leaving ancient shorelines and other pieces of evidence of its lacustrine past visible today. The latest connection with the Mediterranean occurred 12 000 years ago (Çagatay *et al.*, 2009). In addition to this agitated history, the Marmara basin is highly dynamic from a geological point of view, as its Northern slope is traversed by the seismically more active branch of the North Anatolian Fault, the Main Marmara Fault. Gas seeps and water seeps have been unevenly observed in the three major basins found along this fault (from West to East: Tekirdag, the Central basin and Cinarcik) (Géli *et al.*, 2008; Zitter *et al.*, 2008). Gas or cold seeps are enriched in methane and can be seen directly as bubbles coming out from the sediment, or can be detected indirectly by the presence of dark patches revealing areas of strong redox activity in which sulfate reduction occurs just beneath the surface. These patches are often colonized by siboglinid polychaetes and bivalves likely associated with symbiotic bacteria and may be covered by white, little-structured mats of sulfide-oxidizing bacteria. The distribution of gas seeps may provide indications of fault activity (Géli *et al.*, 2008). Water seeps are observed more sporadically. In this case, brackish water trapped in the sediment during ancient lacustrine times (before 14 000 years ago) is expelled through chimneys of authigenic carbonates, as indicated by pore fluid chemistry (Zitter *et al.*, 2008). In addition to the influence of seeps on localized areas, the normal sediments of Marmara, particularly in bays and estuaries, are anoxic and heavily polluted because of incoming currents from the Black Sea, the intense ship traffic and the industrial, agricultural and municipal wastewaters of a densely populated region (mostly Istanbul). A recent study explored the prokaryotic diversity of several sediments in coastal regions based on denaturing gel gradient electrophoresis of amplified SSU rRNA gene fragments, which suggested the dominant

occurrence of fermentative bacteria, denitrifying bacteria and hydrogenotrophic methanogenic archaea (Cetecioglu *et al.*, 2009).

In this study, we present metagenomic data obtained by direct 454 pyrosequencing of DNA from 1000 m-deep plankton of the 0.2–5 μm fraction size (Ma101) and from bottom sediment (1300 m; Ma29) of the Marmara Sea. To relate metagenome data to the more conventional SSU rRNA gene-based diversity, we analyzed in parallel SSU rRNA gene libraries for archaea, bacteria and eukaryotes from the same samples. In addition, we studied the prokaryotic diversity in two other points of the deepest part of the water column (500 and 1250 m depth) below and above the 1000 m-deep point studied in more detail. A comparative metagenomics analysis of Marmara metagenomic data with that of previous studies of deep-sea and surface plankton, subsurface marine sediment, whale carcasses and soil reveals interesting trends.

Materials and methods

Sampling

Samples from the Marmara Sea were retrieved aboard the Atalante in 30 May (sediment) and 6 June (plankton) 2007 during the sampling cruise 'MARNAUT' (http://cdf.u-3mrs.fr/~henry/marmara/index.html). During the cruise, oxygen concentrations at various water columns ranged from 236–238 μmol kg^{-1} in surface waters to 8 μmol kg^{-1} at bottom (clearly suboxic, <10 μmol kg^{-1}; Çagatay *et al.*, 2009). Plankton from the water column in the Central Basin (40° 50.3′N 28°1.4′ E) and from three different depths (samples Ma120 at 500 m, Ma101 at 1000 m and Ma109 at 1250 m, the latter collected 10 m above the bottom) was collected using Niskin bottles mounted on a CTD rosette. Concomitantly, measurements of salinity, temperature and oxygen for each sample were obtained (Supplementary Figure S1). Ma101 (1000 m) had an oxygen concentration of 29.7 μmol kg^{-1}. Its temperature of 14.47 °C and salinity of 38.6 PSU revealed a clear influence of Eastern Mediterranean waters. Seawater filtration was carried out onboard immediately. Total volumes of 5 l (Ma120 and Ma109) and 100 l (Ma101) were filtered through 5 μm-diameter pore filters (TMTP, Millipore, Billerica, MA, USA) and the filtrates passed through 0.22 μm-pore-diameter filters (GTTP, Millipore). Filters were fixed in ethanol and stored at −20 °C. The sediment sample was recovered in a sterile Falcon tube from the surface of a core from a multi-push-core system at 1300 m depth (40° 46.8′N and 29°6.1′ E) and stored at −80 °C.

DNA extraction, PCR amplification, cloning and sequencing

Filters were trimmed into fragments of ca 1 mm^2 with a sterile scalpel and DNA was then extracted using a Mo-Bio PowerSoil DNA extraction kit (Mo-Bio, Carlsbad, CA, USA) following the manufacturer's instructions. DNA from the sediment sample was extracted with the Mo-Bio Ultraclean Soil DNA kit Mega Prep (Mo-Bio) and also purified using the PowerClean DNA Clean-Up kit (Mo Bio) according the manufacturer's instructions.

Archaeal, bacterial and eukaryotic SSU rRNA genes were amplified with at least two different combinations of primers specific for each domain of life: Ar21F (5′-TTCCGGTTGATCCTGCCGGA-3′), ANMEF (5′-GGCTCAGTAACACGTGGA-3′) and the prokaryote-specific primer 1492R (5′-GGTTACCTTG TTACGACTT-3′) for archaea; B27F (5′-AGAGTTTGA TCCTGGCTCAG-3′), B63F (5′-CAGGCCTAACACAT GCAAGTC-3′) and 1492R for bacteria; and EK-42F (5′-CTCAARGAYTAAGCCATGCA-3′), EK-82F (5′-GA AACTGCGAATGGCTC-3′) and EK-1498R (5′-CACCTA CGGAAACCTTGTTA-3′) for eukaryotes. PCR reactions were performed as follows: initial denaturation at 94 °C for 5 min, 32 cycles consisting of a denaturation step at 94 °C for 30 s, annealing at 55 °C for 30 s and extension at 72 °C for 1 min, and followed by a final extension at 72 °C for 7 min. SSU rRNA gene libraries were constructed using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA) and chemically competent TOP10' One Shot cells (Invitrogen). Approximately one hundred clone inserts per domain of organisms and sample were amplified using vector primers and those having the expected size were partially sequenced using the reverse primer (1492R or 1498R, respectively) by Cogenics (Meylan, France). The sequences were deposited in GenBank with accession numbers HM103738—HM103902 (archaeal SSU rDNAs), HM103523—HM103737 (bacterial SSU rDNAs), HM103388—HM103522 (eukaryotic SSU rDNAs) and SRA012098 (454 pyrosequences).

Phylogenetic analysis of SSU rRNA genes and rarefaction analyses

A total of 401 partial prokaryotic SSU rRNA sequences (153, 117, 55 and 76 from samples Ma29, Ma101, Ma109 and Ma120, respectively) were aligned using the NAST alignment tool (DeSantis *et al.*, 2006b) and chimera checked using Bellerophon 3 (http://greengenes.lbl.gov; DeSantis *et al.*, 2006a). The partial high quality sequences were imported in the Greengenes ARB database (236 469 sequences, 18 November 2008 release) and the alignment was manually corrected using the sequence editor included in the ARB software (Ludwig *et al.*, 2004). In the case of eukaryotes, a total of 144 partial sequences (86 for Ma29 and 65 for Ma101) were aligned using the SINA web aligner tool, imported in the SILVA SSURef ARB database (release 100) and manually corrected using ARB (Pruesse *et al.*, 2007). Subsets of sequences were chosen on the basis of neighbor joining trees constructed in ARB. Alignments excluding gaps and ambiguously aligned positions were exported

Table 1 Characteristics of the original metagenomic libraries and of the subsequently selected data used for our comparative metagenomic study.

	Ma29 Marmara deep-sea sediment (1300 m)	*Ma101 Marmara deep-sea plankton (1000 m)*	*ALOHA deep-sea plankton (500 m, 770 m and 4000 m)*	*ALOHA euphotic zone plankton (70 m)*	*Peru sediment mix (1, 16, 32 and 50 m bsf; seafloor 150.5 m)*	*Whale carcasse mix (rib, bone and mat)*	*Waseca soil*
Sequencing method	*Pyrosequencing*	*Pyrosequencing*	*Sanger, fosmid ends*	*Pyrosequencing*	*Pyrosequencing*	*Sanger, shotgun*	*Sanger, shotgun*
Original average size (bp)	181	181	~900	110	100	3400	2400
Size range (bp)	135–240	135–240	—	90–130	90–130	—	—
Average size after trimming (bp)	191.44	191.47	192	110	109	192	192
No of fragments	178 199	183 631	149 022	319 477	266 259	200 722	192 639
Total nucleotides (bp)	34 114 417	35 159 828	28 612 224	35 142 470	29 022 231	38 538 624	36 986 688
GC content (%)	55.63	44.41	51.75	35.35	45.95	47.44	58.03
tRNA (total matches)	187	407	278	286	150	211	187
% COG matches (from no of reads)	28.81	37.45	29.46	22.22	12.82	38.83	30.56
% KEGG matches (from no of reads)	15.62	24.29	17.79	25.70	10.03	21.47	15.54
% of matches against nr (only Ma29 and Ma101)	44.9	57.2	ND	ND	ND	ND	ND
No of rRNA matches (LSU+SSU)	119	476	154	723	146	921	102
COG marker protein matches/Mbp	24.65	50.83	32.52	79.13	46.82	36.72	19.65
Effective genome size (Mbp)	3.66	1.78	2.77	1.59	2.92	2.45	4.59
% of 'ghost reads'	11.8	9.1	0.3	10.0	12.8	0.6	4.6
Reference	This study	This study	DeLong *et al.* (2006)	Frias-Lopez *et al.* (2008)	Biddle *et al.* (2008)	Tringe *et al.* (2005)	Tringe *et al.* (2005)

Abbreviations: COG, clusters of orthologous groups; KEGG, Kyoto Encyclopedia of Genes and Genomes; ND, not determined; rRNA, ribosomal RNA; tRNA, transfer RNA; bsf, below sea floor.

from ARB to reconstruct phylogenetic trees by maximum likelihood using TREEFINDER applying a general time reversible model of sequence evolution (Jobb *et al.*, 2004). Maximum likelihood bootstrap values were inferred using 1000 replicates. The phylogenetic trees were visualized with the program TREEVIEW (Page, 1996). Distance matrices were generated in ARB using previously generated alignments only from good quality sequences and exported in Phylip format. SSU rDNA accumulation curves were obtained for archaea, bacteria and eukaryotes from these matrices at different sequence identity levels (100%, 97%, 95%, 90% and 80%) using DOTUR (Schloss and Handelsman, 2005).

Pyrosequencing of Marmara DNA samples and quality trimming of metagenomic sequences for comparative analyses

DNA from the 0.2–5-μm size 1000-m-deep plankton fraction, Ma101, and from the sediment sample Ma29 was sequenced in $2 \times 1/2$ GS-FLX run (one sample per region; Cogenics, Meylan, France). As a minimum of 5 μg of DNA was needed for pyrosequencing and because of the relatively low amount of DNA obtained from Ma101 (3 μg), this sample was subjected to a whole genome amplification step (Cogenics) to double the DNA amount. The potential bias in DNA amplification, if any, was expected to be low. The total number of valid reads obtained was 545 585. With an average of 182 bp read length, the number of bases cumulated at 99 Mb. We selected five other metagenomic analyses from marine origin (surface and deep-sea waters, ocean subsurface sediments, whale carcasses) and soil (Table 1) to carry out a comparative analysis. As these metagenomic data were generated using different strategies and sequencing technologies and to minimize potential biases owing to differences in sequence length, sequence reads were treated as follows. Ma101 and Ma29 pyrosequence reads were trimmed to the size range of 130–240 bp (average of 192 bp). Surface plankton and Peru margin subseafloor metagenomes consisting of shorter reads were trimmed to a size range of 90–130 bp. Metagenome sequences of whale carcasses and Waseca soil, which were generated by shotgun sequencing, were trimmed differently to get homogenous and comparable sequence fragments. First, only sequences in the size range of 400–3500 bp were retained. Then, 200 nucleotides (soil) and 100 nucleotides (whale carcass) located at 5′ and 3′ of

each read were removed, as they often contained stretches of single nucleotides (low quality sequence) and vector sequences. In addition, reads containing stretches of more than 10 A, T, G, C or N and reads containing strong matches to the NCBI (http://www.ncbi.nlm.nih.org) vector database (manually inspected) were eliminated from our analysis. Sequence length was then homogenized to 192 bp to have a size similar to that of our sequences and, subsequently, duplicate reads were removed. These trimmed sequences were randomly numbered and, from them, sequence fragments were randomly chosen for subsequent metagenome comparison analysis. We used a roughly comparable total amount of sequence from the different metagenomes in our analyses, the smaller metagenomes being the limiting factor (Table 1). The sequences from whale carcasses retained represent a mix of equal proportions of rib, bone and mat whale fall sequences. Fosmid-end sequences from three different depths (500, 770 and 4000 m) of the ALOHA meso- and bathypelagic plankton metagenome were trimmed to 192 bp sizes and vector-contaminated reads were eliminated (three in total). To keep a comparable size of the ALOHA deep-sea plankton with the other chosen metagenomes, we pooled all those trimmed sequences in a single 'ALOHA deep-sea plankton' metagenome (41 995, 55 312 and 52 150 sequences from HF500, HF770 and HF4000, respectively; that is, a total of 149 457 sequences).

Identification and elimination of 'ghost reads'
During preliminary analysis of Marmara pyrosequences, we detected the presence of 9.1–12.8% of artefactual repeated sequences ('ghost reads') in our samples, as well as in other (pyrosequence) metagenomes analyzed in this work (see below), a phenomenon also recently observed by others (Gomez-Alvarez *et al.*, 2009). 'Ghost reads' were also observed in nonpyrosequence metagenomes, but to a lower extent (0.3–4.6%; Table 1). To identify and eliminate 'ghost reads' from our samples, we performed a distribution analysis by plotting, for each metagenome, the number of duplicated reads as a function of duplication size at positions 1, 5, 20, 50 and 80 nt from the start of sequence. As evidenced by the plots (Supplementary Figure S2), a plateau for the number of duplicates is reached in each case when duplication size exceeds ~10 nt, at any start position examined, and levels further for duplication sizes as long as 400 nt. As metagenome sequences are supposed to be randomly distributed, it is highly unlikely that reads start with identical stretches of nucleotides. After additional manual inspection, we determined the following duplication size cut-offs: for the ALOHA deep-sea plankton fosmid-end sequences, the first 20 nt, and for the other metagenomes, the first 15 nt. We then eliminated all the sequences that were identical with respect to these criteria to assure the uniqueness of the reads used for subsequent comparative analyses.

Taxonomic profiling based on rRNA and Clusters of orthologous group (COG) marker protein matches
Trimmed pyrosequences were blasted against the curated rRNA database from Urich *et al.* (2008) containing SSU and large subunit (LSU) rRNA sequences, as well as additional taxonomic identifiers for groups of environmental sequences. BLASTN analysis was performed against SSU and LSU rRNA databases separately. The BLAST output file was used as input file for MEGAN (Huson *et al.*, 2007). The taxonomic affiliation of the sequences was assigned using MEGAN with a cut-off bit score of >90 for the long (Ma101, Ma29, ALOHA deep-sea plankton, whale carcass and soil) and >55 for the short (plankton surface and Peru subseafloor) metagenomic sequences. Reads matching with low bit score values were manually checked. In addition, BLASTX (Altschul *et al.*, 1997) analysis was performed against a database consisting of 31 marker protein families that provide sufficient phylogenetic information to perform taxonomic profiling deriving from 191 reference species with sequenced genomes (Ciccarelli *et al.*, 2006; von Mering *et al.*, 2007). This database was complemented with marker protein sequences from seven recently sequenced genomes that are representatives of additional phyla or groups (*Cenarchaeum symbiosum*, *Nitrosopumilus maritimus* SCM1, *Candidatus Pelagibacter ubique* HTCC1062, *Pseudoalteromonas haloplanktis* TAC125, *Sulfurimonas denitrificans* DSM 1251, *Rhodopirellula baltica* SH 1, *Geobacter uraniireducens* Rf4, *Alteromonas macleodii* 'Deep ecotype'). Only best matches from BLAST analysis were retained and the taxonomic affiliation was determined using MEGAN (original BLAST cut-off e-05, bit score >40, average bit score 69.7).

Estimates of effective genome sizes (EGS)
We applied the *in silico* method developed by Raes *et al* (2007) to estimate the average genome size of microbial communities using metagenomic sequence data.

COG, KEGG and SEED subsystem clustering and cross comparison of metagenomes
The seven selected metagenomes were analyzed by BLASTX against the COG (Tatusov *et al.*, 2003) and KEGG (Kyoto Encyclopedia of Genes and Genomes; Kanehisa *et al.*, 2004) databases applying a cut-off bit score of >40 and normalized to the number of genomic fragments for each metagenome. The number of hits to the functional categories from COG and KEGG with at least 10 and 1 matches, respectively, in all seven metagenomes were normalized to the number of matches in each metagenome

and used for cluster analysis (831 COG functional categories, 148 KEGG functional groups). The percentage of each COG and KEGG functional category shared in the seven metagenomes was determined and shown in cross-comparative heatmaps. The COG functional categories showing the highest variations among the seven metagenomes were determined applying the following cut-off criteria: >10% standard deviation among the metagenomes within the functional categories. The remaining 74 COG functional categories are shown (see below, Figure 5). The seven metagenomes had matches with a total of 207 different KEGG functional groups. KEGG groups without matches in one of the metagenomes and <20 matches in at least another metagenome were eliminated. The remaining 148 functional KEGG groups were used for cluster analysis. Unique, overrepresented and underrepresented groups were identified using s.d. cut-off of 8% and additional manual inspection. The application of different matrices and different sampling depths did not change the metagenome sample tree topology in KEGG analysis (applied to 148 KEGG groups). Hierarchical cluster analysis was performed with MeV 4.4 (Saeed *et al.*, 2003) applying Euclidean distance and Kendall's tau matrices and average linkage clustering.

Screening of functional key enzymes
We constructed reference databases for a number of key enzymes that are diagnostic, or strongly suggestive, of particular metabolic pathways. Only protein sequences with confirmed activity or/and strong similarity to them were included in the database. The metagenomes were blasted against these reference databases and matches showing E-values smaller than 1e-05 were retained. The remaining matches were blasted against nr and the alignments manually inspected to decide about their classification.

Genome recruitment
Genome fragment recruitment analysis was performed using promer and nucmer packages included in the MUMmer 3.0 sequence alignment software tool (Kurtz *et al.*, 2004) Metagenome reads were aligned to the chosen reference genome using the promer and/or nucmer alignment tools under standard conditions. The output file was parsed by show-coords (included in MUMmer 3.0) applying options that knockout overlapping alignments from another reading frame (-k), display the sequence lengths (-l), display start and stop of matching region on nucleotide level (standard), sort the output files according to the reference file (-r) and indicate identities and similarities of each match (default). Nucleotide regions matching the reference genome were identified, and duplicates corresponding to overlapping fragment matches eliminated. The remaining matches correspond to unique nucleotide regions that were nonredundantly covered by the metagenome reads based on amino acid alignments. To avoid the inclusion of non-coding regions as intergenic regions, as true matches, only protein-encoding regions were considered as reference genome length, resulting in the exclusion of ribosomal RNAs and transfer RNAs.

Shared matches and metagenome cross-comparison
Each metagenome was analyzed with promer (MUMmer 3.0) against all other metagenomes using the 'maxmatch' parameter. The results were parsed with the show-coords script with −k and −r parameters as described above. Applying these strict parameters, the average amino-acid identity between the matches was 80.5% for match length coverage of 70%. Only reciprocal matches were retained, and the shared matches between the metagenomes counted. The seven metagenomes were blasted (TBLASTN) against each other using strict-cut-off (bit score >40) and normalization criteria. The numbers of normalized reciprocal best matches were counted and used to construct a distance matrix for subsequent clustering.

Results and discussion

Diversity of Archaea, Bacteria and Eucarya in Marmara deep-sea plankton and sediment based on SSU rDNA gene libraries
To provide a contextual framework of microbial diversity to our metagenomic analyses, we constructed SSU rRNA gene libraries of the same samples plus, in the case of plankton, two additional depths in the aphotic water column, using various combinations of domain-specific primers, followed by SSU rDNA sequencing, BLAST comparison and phylogenetic analysis. We then compared the diversity retrieved in this way with the information derived from Marmara metagenomic data, both by using the identified SSU rRNA gene sequences and taxonomically relevant marker proteins as classified using the NCBI taxonomy by MEGAN (Huson *et al.*, 2007; see below).

A total of 165 archaeal and 215 bacterial SSU rRNA sequences of high quality were generated from gene libraries of the Marmara water column at 500 m (Ma120), 1000 m (Ma101), 1250 m (Ma109) and from a sediment sample (Ma29) in the Marmara Sea bottom. The eukaryotic diversity was studied only in the two samples selected for pyrosequencing, Ma101 and Ma29, with a total of 135 sequences generated. Short and/or poor-quality sequences were discarded from our analyses. As expected, the sediment sample Ma29 harbored a higher diversity of organisms belonging to the three domains of life than plankton samples, which was also reflected in the corresponding rarefaction curves (Supplementary Figure S3), though this trend was more apparent for archaea and bacteria. Figure 1

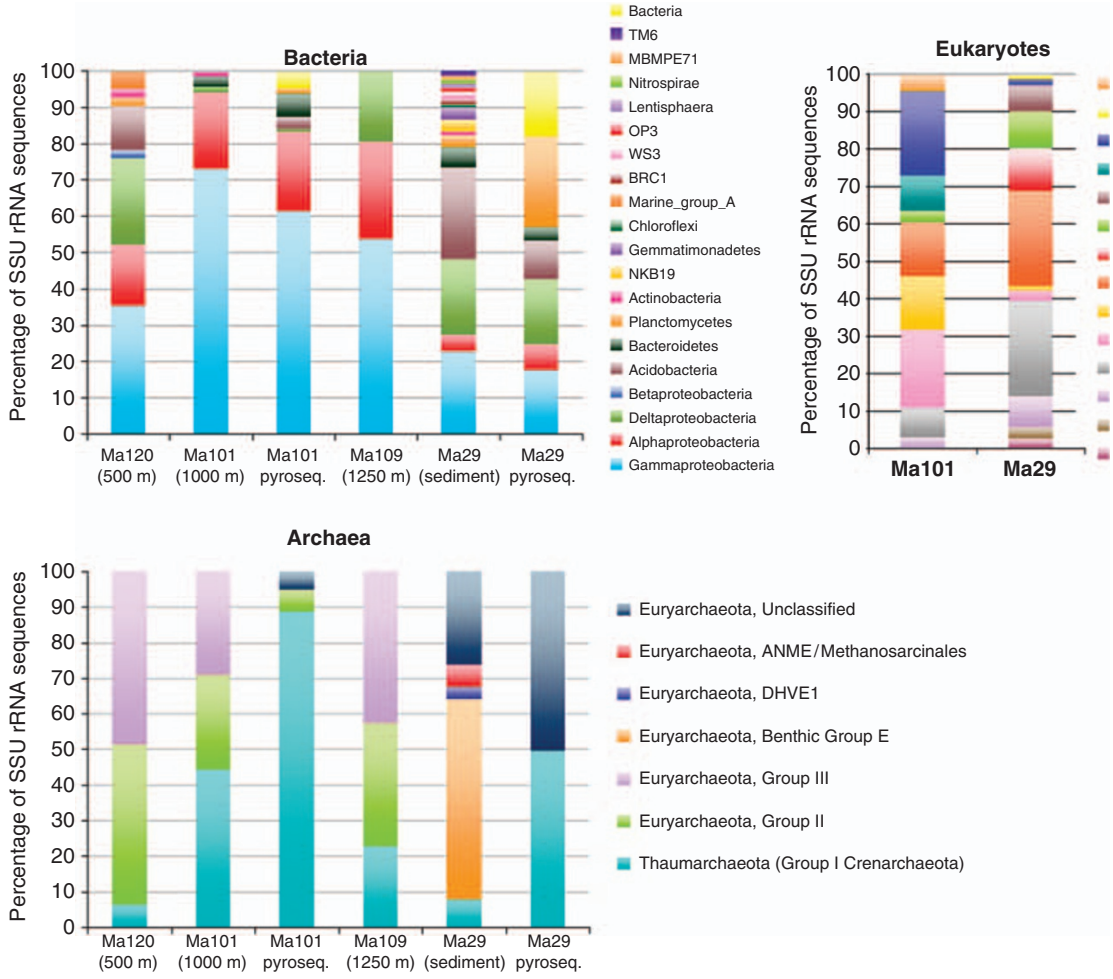

Figure 1 Relative proportions of different archaeal, bacterial and eukaryotic taxa based on SSU rDNA sequences detected in gene libraries and metagenomic data in deep-sea plankton and bottom sediment of the Marmara Sea. Pyrosequence indicates SSU rDNA percentages in metagenomic pyrosequence data as classified by MEGAN; no eukaryotic SSU rDNA sequences were identified in them.

shows the relative proportions of the different taxa identified in the different SSU rDNA libraries as deduced from individual phylogenetic analyses (see Supplementary Figures S4 to S15).

Planktonic bacteria were more diverse in mesopelagic waters (500 m), with seven different bacterial phyla detected by phylogenetic analysis (Figure 1 and Supplementary Figures S4 to S9), than in bathypelagic waters (1000 and 1250 m). Nevertheless, libraries from all three samples were largely dominated by Proteobacteria, Gammaproteobacteria being largely the most represented group at 1000 and 1250 m depth. At 1000 m, 85% of the Gammaproteobacteria affiliated to the common opportunistic marine planktonic genus *Alteromonas* (Figure 1 and Supplementary S4). Alphaproteobacteria represented ~20% of the bacterial population in all plankton libraries, and were dominated by members of the ubiquitous SAR11 group, followed by members of the *Roseobacter* clade (Supplementary Figure S6). The latter is also very abundant in marine environments and occupies a variety of ecological niches showing activities that go from sulfur oxidation and the production of secondary metabolites to carbon monoxide oxidation and aerobic anoxygenic photosynthesis (Wagner-Dobler and Biebl, 2006; Brinkhoff *et al.*, 2008). Deltaproteobacteria were underrepresented at 1000 m, but were the third most represented bacterial group at 500 m and 1250 m (Figure 1). However, whereas in the 500 m library all deltaproteobacterial sequences affiliated to the typical deep-sea planktonic group SAR324 (Wright *et al.*, 1997; López-García *et al.*, 2001), sequences from the 1250 m sample, a suboxic sample at only 10 m above the sea bottom, were related also to deep-sea sediment environmental clones affiliated to the genus *Nitrospina* or were sequences of difficult affiliation within the Deltaproteobacteria (Supplementary Figure S7). Acidobacteria, a phylum that was found to be abundant in deep-sea Mediterranean waters (Quaiser *et al.*, 2008), was also relatively well represented in mesopelagic waters, but not in deeper waters. By contrast, Acidobacteria were one of the three dominant groups in sediment libraries, together with Gamma- and mostly sulfate-reducing Delta-proteobacteria (Figure 1,

Supplementary Figures S7 and S8). Sediment Gammaproteobacteria clustered apart from plankton sequences, being mostly related to clone sequences from deep-sea sediments. One sequence was related to the symbiont of the anaerobic annelid *Olavius algarvensis* (Supplementary Figure S5; Woyke *et al.*, 2006), relatives of which have been found in association with other gutless oligochaete worms in the Mediterranean (Ruehland *et al.*, 2008). This suggests that some of the sediment-associated bacteria form stable endosymbiotic associations with a variety of anaerobic eukaryotic hosts. In addition to Proteobacteria and Acidobacteria, sequences ascribing to another 14 phyla and candidate divisions were identified in the sediment sample Ma29 (Figure 1 and Supplementary Figure S9), highlighting the extremely rich and largely unexplored bacterial diversity of deep marine sediments, similar to previous observations in other Mediterranean sediments (Polymenakou *et al.*, 2005).

Archaea showed lower diversity than bacteria as indicated by rarefaction curves showing that our libraries had practically reached saturation at 97% sequence identity, especially in the bathypelagic samples (Supplementary Figure S3). Although group I Crenarchaeota/Thaumarchaeota (Brochier-Armanet *et al.*, 2008) frequently dominate in deep ocean waters (Karner *et al.*, 2001; DeLong *et al.*, 2006), our planktonic samples seemed dominated by Euryarchaeota based on SSU rDNA sequences (between 57% and 94% of archaeal sequences) as has been observed in other deep-sea localities (Martin-Cuadrado *et al.*, 2008; Figure 1 and Supplementary Figure S10 and S11). This would be in agreement with recent observations that group I archaea in the East Mediterranean meso- and bathypelagic waters may have more variable relative proportions than those suggested by previous studies (Varela *et al.*, 2008). However, these proportions partially reflect the use of one Euryarchaeota-biased primer sets used to amplify archaea ($\sim$50% of sequences), and hence should be taken with caution. Thaumarchaeota reached important levels in bathypelagic samples (Figure 1), but although present in the sediment Ma29, they represented $<$10% of the total archaeal sequences. All the Ma29 thaumarchaeotal sequences associated to the sediment sample clustered with *Cenarchaeum/Nitrosopumilus* or with a group composed of environmental sequences retrieved from deep-sea environments (Lost City, Mariana Trough) and from the substratum of the fish tank where *N. maritimus* was isolated from (Konneke *et al.*, 2005). By contrast, planktonic sequences clustered with other cosmopolitan operational taxonomic units (OTUs) identified in Pacific, Antarctic and Mediterranean meso- and bathypelagic waters (DeLong *et al.*, 2006; Martin-Cuadrado *et al.*, 2008; Supplementary Figure S10). Within the Euryarchaeota, groups II and III were roughly equally represented in the three water column depths studied (Figure 1). Group II Euryarchaeota were only detected in plankton, forming two major lineages (Supplementary Figure S11) that were not related to proteorhodopsin-containing group II Euryarchaeota (Frigaard *et al.*, 2006). The generally underrepresented group III Euryarchaeota was relatively abundant in the deepest part of the water column in Marmara. Euryarchaeota, especially of the benthic group E (Vetriani *et al.*, 1999), dominated the SSU rRNA gene library from the sediment sample. About 25% of the Ma29 archaea were scattered in ill-defined euryarchaeotal groups composed of environmental sequences coming, mostly, from deep-sea sediments or hydrothermal vents. Finally, around 7% of Ma29 archaea belonged to the Methanosarcinales, with some sequences very closely related to members of the ANME2 group (Figure 1 and Supplementary S12). The presence of ANME sequences is not surprising, as the Marmara Sea harbors several scattered cold seep areas (Géli *et al.*, 2008; Zitter *et al.*, 2008) where the anaerobic oxidation of methane likely occurs. However, the relatively low proportion of archaea is consistent with the 'normal' bottom-sediment nature of the Ma29 sediment.

The diversity of microbial eukaryotes was high in both Marmara deep-sea plankton and sediment, as shown by rarefaction curves far from saturation (Figure 1 and Supplementary S3). As expected in waters from this depth, our planktonic SSU rDNA libraries were dominated by alveolate lineages, which accounted for 49.3% of the total clones and comprised sequences belonging to the parasitic marine alveolate groups I (Duboscquellida) and II (Syndiniales), as well as to core dinoflagellates (Supplementary Figure S13). Alveolates were also very abundant in sediment libraries (29.5%), with typical dinoflagellate sequences more abundant than sequences belonging to the marine alveolate groups I and II. One ciliate sequence was also identified in Ma29 (Supplementary Figure S13). The second more abundant group in Ma101 plankton was that of the Radiozoa (Rhizaria) with both representatives of the Acantharea and Polycystinea. Although Acantharea were slightly more abundant, all the sequences belonged to a single phylotype, closely related to an environmental sequence retrieved from Arctic waters (Lovejoy *et al.*, 2006), whereas the Polycystinea showed a higher diversity (Supplementary Figure S14). A few sequences of haptophytes were detected in Ma101, which likely correspond to sinking cells, though it might be possible that some haptophytes are still active at this depth because of their mixotrophic capabilities. A few sequences of fungi and metazoa (ctenophores and copepods) were identified in the planktonic sample as well (Supplementary Figure S15). In the sediment libraries, together with the alveolates, opisthokonts (fungi and metazoa) were the most abundant, accounting for up to 30% of the total sequences. We also detected members of the Amoebozoa, Telonemia, Cercozoa and Stramenopiles.

Several of the Stramenopile (heterokont) sequences corresponded to diatoms, suggesting that diatoms sink while conserving their DNA intact more successfully than other photosynthetic eukaryotes, though we cannot exclude that they corresponded to heterotrophic diatom species, which have recently been found in coastal sediments (Blackburn *et al.*, 2009).

Microbial diversity in Marmara deep-sea plankton and sediment based on SSU rDNA and protein marker gene pyrosequences

We compared the community composition deduced from SSU rDNA libraries to that deduced from the detection of *SSU rRNA* gene fragments in pyrosequence reads followed by their taxonomic classification by MEGAN (Huson *et al.*, 2007). However, this could only be performed for prokaryotic communities, as we detected a too small number of eukaryotic *SSU rRNA* gene fragments (only 5) in Ma101 and Ma29 pyrosequences, and none of them could be classified in a given taxonomic group with certainty. The low number of eukaryotic *rRNA* gene matches was likely because of both, the larger genome sizes of eukaryotes (decreasing the probability of sequencing a particular gene) and the fact that they are quantitatively less abundant than prokaryotic cells in our samples. The relative proportions of the major bacterial groups as classified by MEGAN in Ma101 plankton pyrosequences were comparable to those identified in SSU rDNA libraries (Figure 1) with a large dominance of Gammaproteobacteria (with 73.2% clones, 61.6% rRNA matches) followed by Alphaproteobacteria (21.1% clones, 21.9% rRNA matches) and Bacteroidetes (2.8% clones, 6.2% rRNA matches). The classification of bacterial SSU rDNA clones and pyrosequences was also rather congruent in the sediment sample Ma29, with Gammaproteobacteria, Alphaproteobacteria, Deltaproteobacteria and Bacteroidetes showing similar proportions. On the other hand, Planctomycetes seemed to be underrepresented in the clone library (3.3%) compared with the rRNA matches (25%), whereas an opposite trend was observed for the Acidobacteria (10% in pyrosequence matches, whereas they represented 25.6% in the clone library). These variations were probably due to differences in primer specificities as it is known for Planctomycetes (Vergin *et al.*, 1998), as well as to the statistical error associated to the total low numbers of assigned SSU rRNA matches in the sediment sample. A relative high abundance of Planctomycetes was already reported in Mediterranean deep-sea sediments, as they represented up to 14% (7.9–14%) in SSU rRNA clone libraries constructed from four different sediment samples (Polymenakou *et al.*, 2005). In all, $\sim$20% of the bacterial sequences in our gene libraries affiliated to candidate divisions and little abundant phyla. A similar percentage appeared as unclassified bacteria using MEGAN (Figure 1). By contrast, the affiliation of archaeal pyrosequences by MEGAN did not correlate well with that of SSU rDNAs in libraries, although this may be partly due to primer-induced biases toward the Euryarchaeota. In the plankton sample, Thaumarchaeota pyrosequences dominated (89%), whereas they accounted for 43% in gene libraries. In the sediment sample Ma29 only six archaeal SSU rRNA matches were identified in the sediment metagenome, which made the comparison with SSU rDNA data unreliable.

In addition, we carried out BLASTX analyses of Ma101 and Ma29 pyrosequences against the nr database in GenBank. The output file was then used as input in MEGAN to classify the sequences according to its taxonomic identifiers. From the totality of raw sequences, 44.9% (Ma29) and 57.2% (Ma101) had matches in the nr database (Table 1). Archaeal matches accounted for 10.3% (95% of which belonged to the Thaumarchaeota) and 2.32% (47% Thaumarchaeota) in Ma101 and Ma29, respectively. Although these values compare well with the respective percentages of archaeal SSU rDNA matches in pyrosequences, they are potentially highly biased because of the lack of complete genome sequences of group II and III Euryarchaeota in databases.

Microbial diversity in Marmara deep-sea plankton and sediment compared with other metagenomic data sets

We compared the Marmara sequences against a collection of selected published metagenomes that were adequately trimmed and/or merged to minimize potential biases because of differences in sequence length and read depth (Table 1, see Material and methods). We selected all the metagenomes available at the time of analysis corresponding to deep-sea plankton that were of sufficient size to be compared with our data sets in addition to sensible outgroup metagenomes. These included the data sets of deep-sea plankton fosmid-end sequences in the Pacific ALOHA water column (500–4000 m depth; DeLong *et al.*, 2006) and the metagenomic sequences of surface plankton in the same water column (Frias-Lopez *et al.*, 2008). For marine sediments, the only metagenomic data available corresponded to Peru margin subsurface samples (Biddle *et al.*, 2008) that, although not very deep (150 m depth), could still be a good representative of this type of habitat. We included the Waseca soil metagenome (Tringe *et al.*, 2005), as soils correspond to the ecological compartment where the ultimate degradation of organic matter takes place on continents, as is the case of sediments in oceans. Finally, we also included whale carcass metagenomic data (Tringe *et al.*, 2005), because of both its marine deep origin and its transitional location between bottom sediment and deep plankton. The microbial diversity captured in the DNA sequences from the seven different metagenomes was analyzed using rRNA and conserved marker

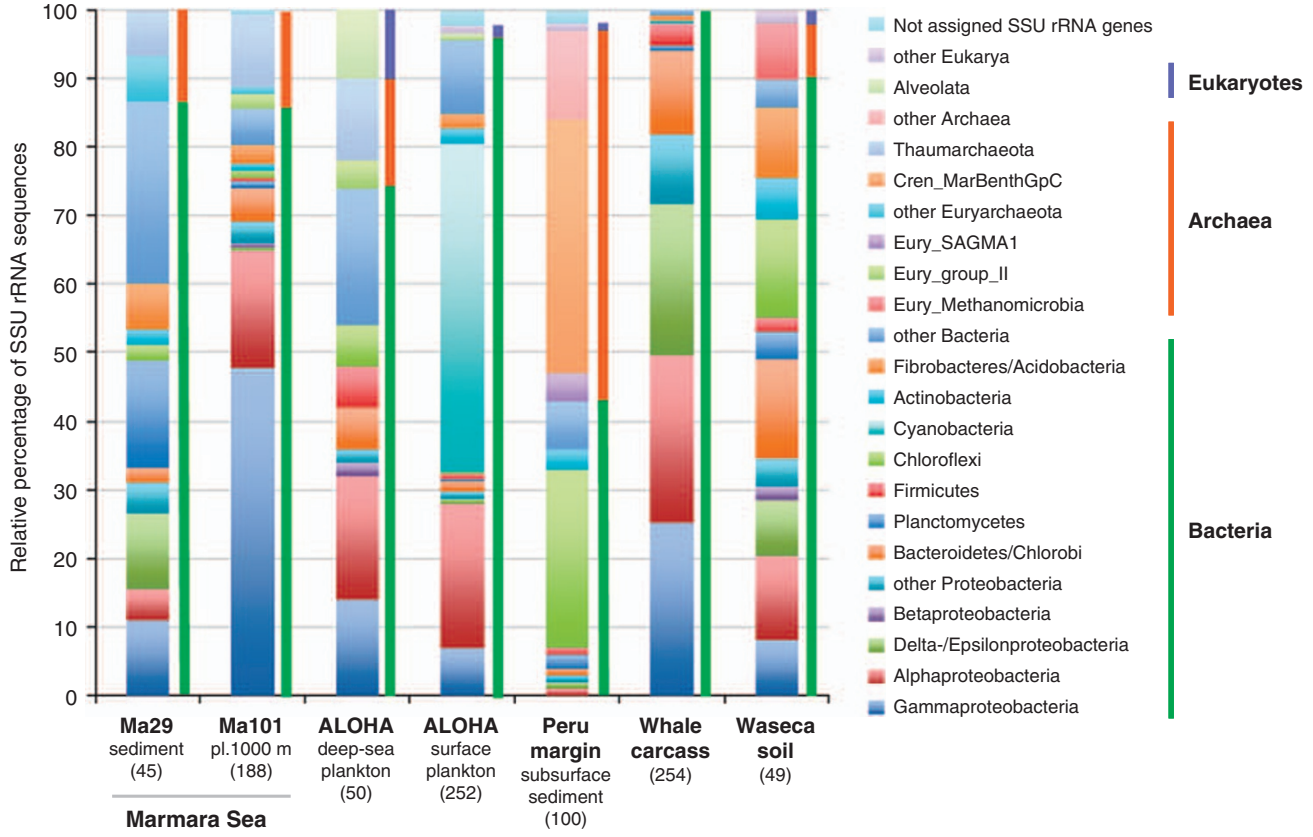

Figure 2 Relative proportions of different archaeal, bacterial and eukaryotic taxa identified in different metagenomic data sets based on their content of SSU rDNA sequences. The taxonomic assignment was carried out using MEGAN. The total number of SSU rDNA matches in each metagenome is indicated in parentheses.

protein sequence matches. Metagenome sequences were blasted (BLASTN) against a curated SSU and (LSU) database containing detailed taxonomic identifiers including rRNA genes from environmental samples (Urich *et al.*, 2008). They were also blasted (BLASTX) against 31 families of marker proteins from 198 reference species covering 36 different taxonomic groups of all three domains (18 for Bacteria, 10 for Archaea and 8 for Eucarya) that generally provide sufficient phylogenetic information as to perform taxonomic profiling (Ciccarelli *et al.*, 2006). Although this approach suffers from some bias, given that many of the phylotypes identified by SSU rRNA sequences belong to lineages for which genome sequences are not yet available, the overall taxonomic profile based on these 31 protein markers corroborated that obtained from SSU rDNA sequences, although with some minor quantitative differences (data not shown).

The taxonomic classification of the rRNA and marker protein matches in the different metagenomes was analyzed using MEGAN. As the metagenome sequences are too short for detailed affiliation to taxonomic groups, we limited the assignment to the phylum level. We first identified rRNA matches to known SSU rRNA sequences. The affiliation to SSU rRNA sequences excluding the LSU rRNA matches represents an adequate way to capture the largest diversity, as SSU rRNA sequences are available from sequenced genomes, but also from environmental species only known by their SSU rRNA genes. The highest number of SSU rRNA hits was found in the plankton surface (272) followed by the whale carcass metagenomes (254) and Ma101 (188) and ALOHA deep-sea fosmid ends (50; Table 1). The lowest number of rRNA hits occurred in sediment Ma29 (45) followed by the soil metagenome (49) and Peru subseafloor (100), likely revealing a larger genome average size (see below).

At the level of domains, bacteria were the most abundant group in all but the Peru metagenome (Figure 2). In the Peru subseafloor sediment, 54% of the rRNA matches affiliated to Archaea (37.5% of total affiliated to Marine benthic group C, Crenarchaeota), although archaeal rRNA matches were only detected at 16, 32 and 50 m, but not at 1 m below the seafloor, as already reported (Biddle *et al.*, 2008). A relatively high proportion of archaeal rRNAs was found in both planktonic deep-sea metagenomes, Ma101 and ALOHA deep-sea sequences accounting for 13.6% and 16.0% (Thaumarchaeota: 11.1% and 12.0%), respectively. In the sediment sample as well as in soil, archaeal rRNA matches accounted for 13.3% and 8.2%, respectively. Surprisingly, no archaeal rRNA sequence was detected in the whale carcass metagenome.

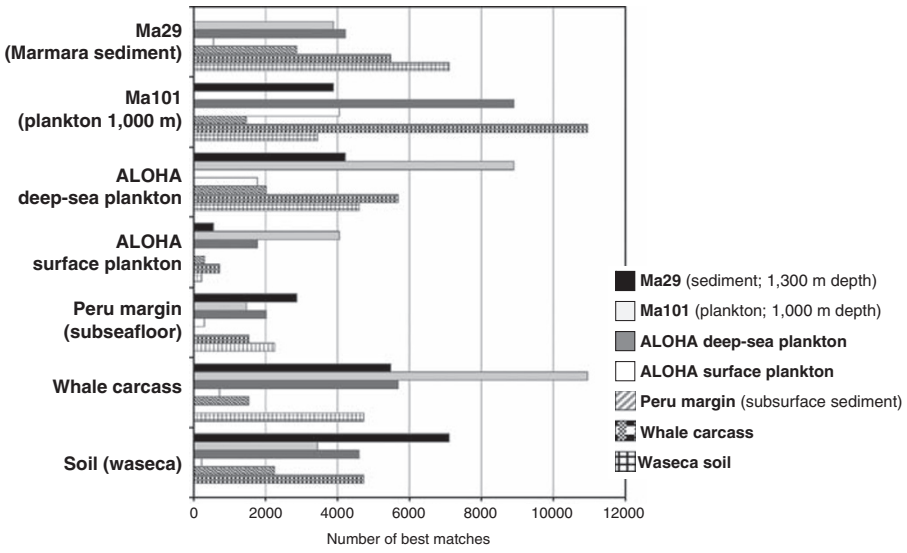

Figure 3 Best reciprocal TBLASTN high scoring pairs (HSPs) between selected metagenomes.

Within bacteria, Gammaproteobacteria were dominant (47.9%) in the Ma101 plankton sample and they were relatively abundant in the whale carcass metagenome (25.2%), but were much less abundant in other metagenomes, with values around 10% (Figure 2). The proportion of Alphaproteobacteria was much more stable in all plankton and the whale carcass metagenomes, with values around 20%. The Marmara sediment and the soil sample had a relatively similar distribution of rRNA hits in taxonomic groups. The two more different taxonomic profiles corresponded to the ALOHA surface plankton metagenome, which was dominated by cyanobacteria (47.8%) and to the subseafloor ecosystem (Peru margin) where archaea dominated followed by bacteria belonging to the Chloroflexi (26.0%; Figure 2). Eukaryotic SSU rDNAs were scarce with no or very few matches detected (<5%); only in the ALOHA deep-sea plankton metagenome they represented around 10% of the total SSU rDNAs detected. Taking into account that eukaryotic rRNA genes are usually organized in large tandems with, sometimes, hundreds of gene copies, these observations reinforce the idea that, although diverse, eukaryotes are minor components of the microbial diversity, at least in terms of cell numbers.

Comparison of Marmara deep-sea plankton and sediment with other metagenomes

We then compared seven selected metagenomic data sets by reciprocal TBLASTN analyses. We identified high scoring sequence pairs and computed the number of reciprocal best matches. As can be seen in Figure 3, the highest number of best reciprocal matches occurred in the Marmara deep-sea plankton Ma101 and whale carcass metagenomes, followed by Ma101 and the ALOHA deep-sea plankton data set. The Peru margin data set accounted for the lowest number of high scoring pairs with the rest of the metagenomes, followed by the ALOHA surface sample.

These results were further confirmed by the analysis of shared sequence reads between metagenomes. For this, each metagenome was analyzed against all other metagenomes using the promer alignment tool of MUMMER (Kurtz *et al.*, 2004) with strict parameters and considering only reciprocal best matches. Under these conditions, only highly conserved proteins that are present in the metagenomes can be identified. The highest number of reciprocal matches was again identified between the whale carcass and the Marmara deep-sea plankton metagenome (Ma101) with 9666 matches followed by the Ma101/ALOHA deep-sea data set with 8367 matches (Figure 4a). As the number of reads among these three metagenomes differs to some extent for the ALOHA deep-sea metagenome (149 022) compared with Ma101 (183 621) and whale carcass (200 722), the shared reads to the ALOHA deep-sea data set represent a minimum value of shared reads. Nevertheless, as the number of shared reads with the other metagenomes is much lower, these three metagenomes are clearly the most similar, sharing at least 1859 reads. Most of those shared reads encoded housekeeping proteins as ribosomal proteins and transfer RNA synthetases, corresponding essentially to the marker proteins mentioned above. Cluster analysis using the number of shared reads, did indeed group together those three metagenomes (Figure 4b). The Marmara sediment Ma29 clustered with the soil sample and all these five metagenomes were closer among them than with the Peru subseafloor and the ALOHA surface water metagenomes.

This clustering was further confirmed by the comparison of COG and KEGG categories in the different data sets (see below).

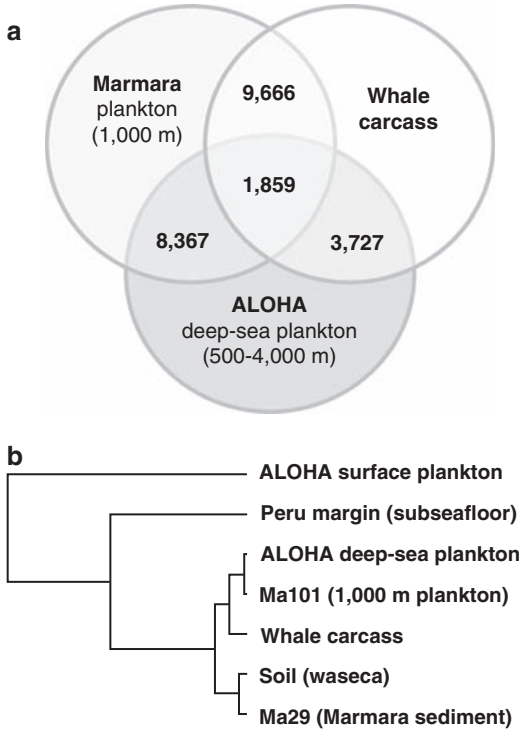

Figure 4 MUMMER-based identification of shared sequences in metagenomes. (**a**) Number of shared sequences in Marmara and ALOHA deep-sea plankton and whale carcass metagenomes. (**b**) Cluster analysis of selected metagenomes based on shared sequences.

COG and KEGG functional categories

The predicted protein sequences of the seven metagenomes were compared against different databases including COGs (Tatusov *et al.*, 2003), KEGG (Kanehisa *et al.*, 2004) and SEED subsystems (Overbeek *et al.*, 2005). In all, but the plankton-surface and the subseafloor metagenomes (both encompassing short reads, ~100 bp) the proportion of matches to the COG database exceeded largely the KEGG and SEED matches (Supplementary Figure S17). The sequences were categorized and the presence of functional protein groups and metabolic traits compared among the different metagenomes using cluster analysis. At the level of COG categories, only slight variations between the metagenomes could be observed. The highest variations corresponded to the categories of signal transduction (category T) and translation and ribosome biogenesis (category J) and, more precisely, to the ratio between the two categories (Supplementary Figure S18). Notably, other housekeeping genes, such as those involved in transcription (category K) seemed more stable across metagenomes than those involved in translation. The ratio of genes involved in signal transduction versus those involved in translation was lowest in the surface ALOHA metagenome and in the Peru subsurface metagenome, whereas it was intermediate in the deep ALOHA and Marmara plankton and much higher in the rest of the metagenomes. The fact that genes involved in signal transduction were more relatively abundant in the whale carcass, in Marmara sediment and in soil metagenomes is in agreement with the idea that the microbial communities in these structured biotopes entertain more complex interactions that involve cell-to-cell communication. Similarly, the higher ratio of signal transduction versus translation genes in deep-sea plankton compared with that in surface waters is suggestive of a higher cell-to-cell interaction level in deep-sea planktonic communities. This would be in agreement with the idea that deep-sea planktonic life occurs mostly associated to particles of sinking organic matter, where biofilms may form (Martin-Cuadrado *et al.*, 2007). Finally, the low ratio of signal transduction versus translation genes in the Peru subsurface is likely explained by the dominance of this particular sample by Crenarchaeota, generally characterized by compact genomes, and by the particular stable environmental constraints of this extreme habitat that result in a genetically distinct community (Biddle *et al.*, 2008).

A hierarchical cluster analysis of the different COG categories (831 COGs having at least 1 match in each of the 7 metagenomes) showed that the two deep-sea plankton samples plus the whale carcass metagenomes, on the one hand, and the soil and Ma29 sediment, on the other hand, grouped together (Figure 4 and Supplementary Figure S19). To identify functional categories that might be specific for a metagenome we determined 74 COGs that showed the highest variations applying a cut-off standard deviation of 10% within the functional group followed by additional manual inspection (Figure 4). The more highly represented COGs in the deep-sea plankton sample (Ma101) corresponded to a chemotaxis protein involved in environmental sensing and cell response, and the 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase involved in the biosynthesis of aromatic amino acids (Table S1). The sediment sample (Ma29) was largely enriched in a variety of COGs, many of which were related to anaerobic and/or autotrophic metabolisms and to sulfur cycling (Table S1), which is in good agreement with microbial communities associated to anoxic sediments.

The hierarchical cluster analysis of KEGG functional categories yielded a similar result to that observed using COG categories, with the deep planktonic and the whale carcass and the soil and sediment metagenomes, respectively, clustering together (Figure 4 and Supplementary Figure S20). The deep-sea plankton sample Ma101 showed relatively high levels of sequences affiliated to bacterial chemotaxis, flagellar assembly, arachidonic acid metabolism, type III secretion system and gamma-hexachlorocyclohexane degradation. In Ma29, the most abundant genes corresponded to electron transfer carriers, type II secretion system, signal transduction mechanisms and energy metabolism.

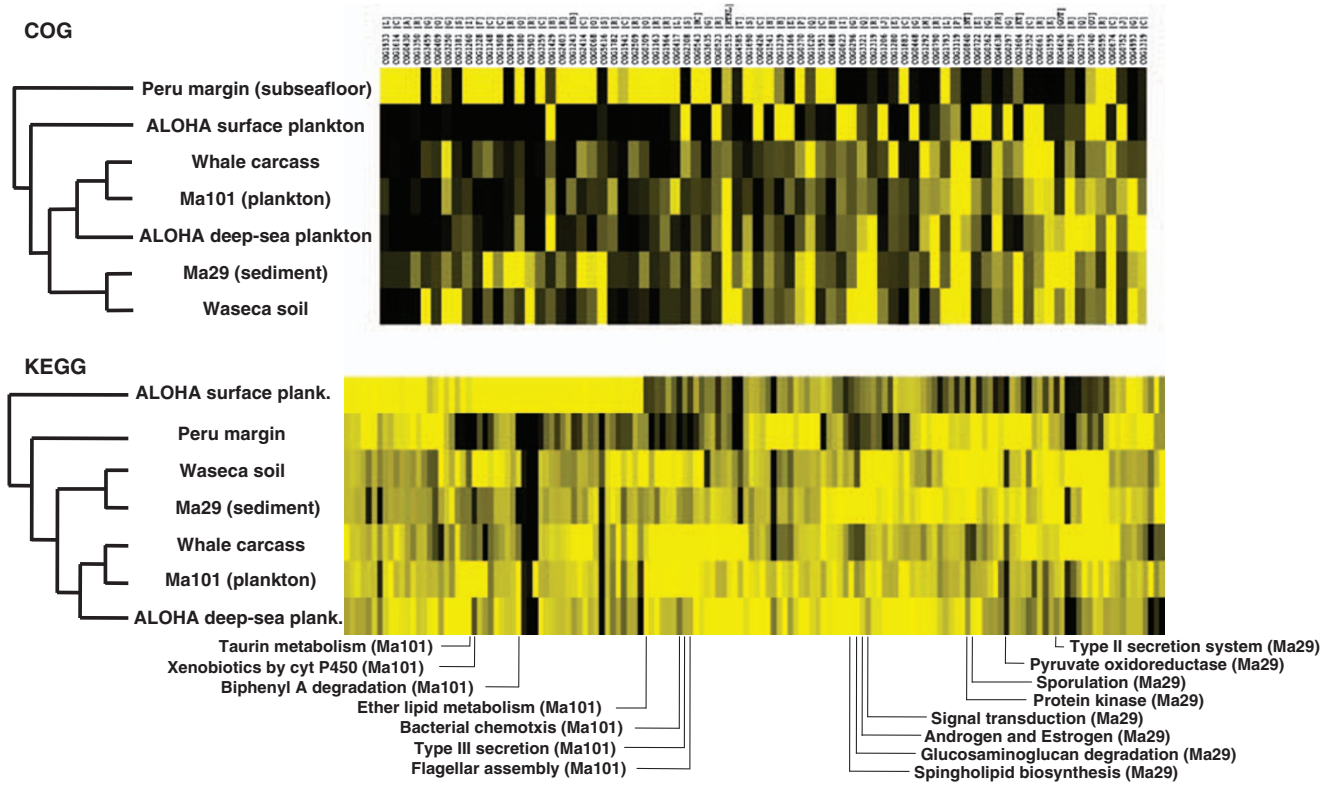

Figure 5 Cluster analysis of several metagenomes based on matches to different COG and KEGG categories normalized to the respective number of sequence fragments. For COGs, only functional categories having at least 10 matches in all 7 metagenomes were retained (831 COGs), from which the most frequent 74 are shown (see also Supplementary Figure S19). For KEGGs, only functional categories having at least 1 match in all 7 metagenomes or at least 20 matches in 1 metagenome were retained (148 KEGGs; see also Supplementary Figure S20). The names of particular KEGG categories showing strong differences across metagenomes are shown at the bottom.

Estimation of the EGS based on COG marker proteins

Recently, Raes *et al*. (2007) established an *in silico* method to estimate the average genome size of microbial communities using metagenomic sequence data. This method is based on the inverse relationship of the genome size and the marker gene density and includes the normalization to the number of sequence reads and the total number of base pairs of the analyzed metagenome. The selected marker genes encode for 35 COG marker proteins present in all genomes with low and constant copy numbers, independently of genome size. This correlation was identified by simulations using fragmented complete genome sequences and applied to metagenomic sequences with fragment lengths of 400–1100 bp and a bit score cut-off >60 for the identification of marker protein matches by BLAST analysis. To test the applicability of this correlation to shorter reads and to get an idea of the average genome size of microorganisms in the deep-sea plankton and sediment, we applied the described function to our data. The 35 COG marker proteins were blasted against all 7 metagenomes. As the bit score in BLAST analysis depends on the aligned fragment length, we first evaluated different bit score cut-offs ranging from 20 to 80. We observed a strong variability in the number of marker protein matches and, consequently, in the EGS estimations using bit score cut-offs in the range of 60 and above. This indicates that this approach strongly depends on fragment length. Hence, it can be at most indicative when applied to fragments with average lengths between 100 and 200 bp. Nevertheless, when applying a bit score cut-off of >40 for the detection of marker genes, we obtained an EGS of 3.66 Mbp for the sediment community (Ma29) and 1.78 Mbp for the deep-sea plankton (Ma101; Table 1 and Supplementary Figure S21). Using the untrimmed original metagenome fragments from one GOS surface plankton sample and the Waseca soil ecosystem and applying a bit score -off of >60, Raes *et al*. (2007) obtained average prokaryotic EGS of 1.6 Mbp and 4.74 Mbp, respectively. These values are comparable with our results obtained from the ALOHA surface plankton pyrosequences (1.71 Mbp) and the trimmed Waseca soil fragments (4.59 Mbp) using a bit score cut-off threshold >40. However, although Raes *et al*. (2007) estimated the whale fall average genome size to 3.55 Mbp using the average of all three different whale carcasses samples, our estimation yielded only 2.45 Mbp (cut-off >40). This might be partly because of differences in the actual sequences considered for the estimation, as in our study we used an equal mixture of the different whale ecosystems trimmed to shorter sequence lengths. At any rate, despite the

variability of the EGS values obtained depending on the thresholds used, these estimates are consistent with smaller genome sizes associated to plankton samples and larger genome sizes associated to soil and sediment samples, with the notable exception of the highly divergent and archaea-dominated Peru subsurface.

Functional key enzymes

In addition to the general overview on the community functional potential that the relative abundance of functional COG and KEGG categories provide from the different metagenomes, we looked for particular proteins that are considered to be diagnostic for particular enzymatic pathways and, hence, for particular metabolic capabilities. These included ammonia monooxygenase AmoA, AmoB and AmoC subunits (nitrification), 4-hydroxybutyryl dehydratase (CO_2 fixation by the 3-hydroxypropionate/4-hydroxybutyrate pathway), dissimilatory sulfite reductase DsrA and DsrB subunits (sulfate respiration), dissimilatory nitrite reductase subunits NirK and NirS (nitrate respiration), nitrogenase subunits NifH and NifD (nitrogen fixation), carbon-monoxide dehydrogenase CoxLMS subunits (CO oxidation), RuBisCO (CO_2 fixation), sulphatase (degradation of sulfonated heteropolysaccharides), hydroxylamine oxidoreductase HAO (anammox), methyl coenzyme A reductase (anaerobic oxidation of methane) and C-P lyase (phosphonate utilization; Figure 6 and Supplementary Table S2).

The ammonium monooxygenase is the key enzyme in the first step of nitrification. Thaumarchaeota are major factors in the oxidation of ammonia to nitrite in soil and oceans as suggested by the dominance of archaeal over bacterial *amoA* genes (Leininger *et al.*, 2006; Yakimov *et al.*, 2007). However, not all deep-sea Thaumarchaeota possess *amoA*, suggesting that many deep-sea archaea are not chemolithoautotrophic ammonia oxidizers (Agogue *et al.*, 2008). In our case, archaeal *amo* genes were fairly abundant in the deep-sea plankton sample Ma101 (32 matches) as compared with the rest of metagenomes (below 1–2 matches), whereas bacterial *amo* genes were detected in the Ma101 metagenome in much lower numbers (1 match). This suggests that bathypelagic Thaumarchaeota in the Marmara Sea are indeed ammonia oxidizers, whereas those in the open Pacific Ocean are likely not (or to much lesser extent). In addition, Ma101 Thaumarchaeota seem to be chemolithoautotrophic, as we identified a high number of matches with strong identities and bit scores (ranging from 58–151) to 4-hydroxybutyryl dehydratase, a key enzyme in the 3-hydroxypropionate/4-hydroxybutyrate pathway for autotrophic CO_2 fixation in group I archaea (Berg *et al.*, 2007). In addition to this C fixation pathway, we identified a number of hits to the RuBisCO large subunit (rbcL), indicating the presence of the more conventional Calvin cycle for CO_2 fixation in the different metagenomes (Supplementary Table S2).

We identified genes involved in nitrate respiration in the whale carcass, Marmara sediment and

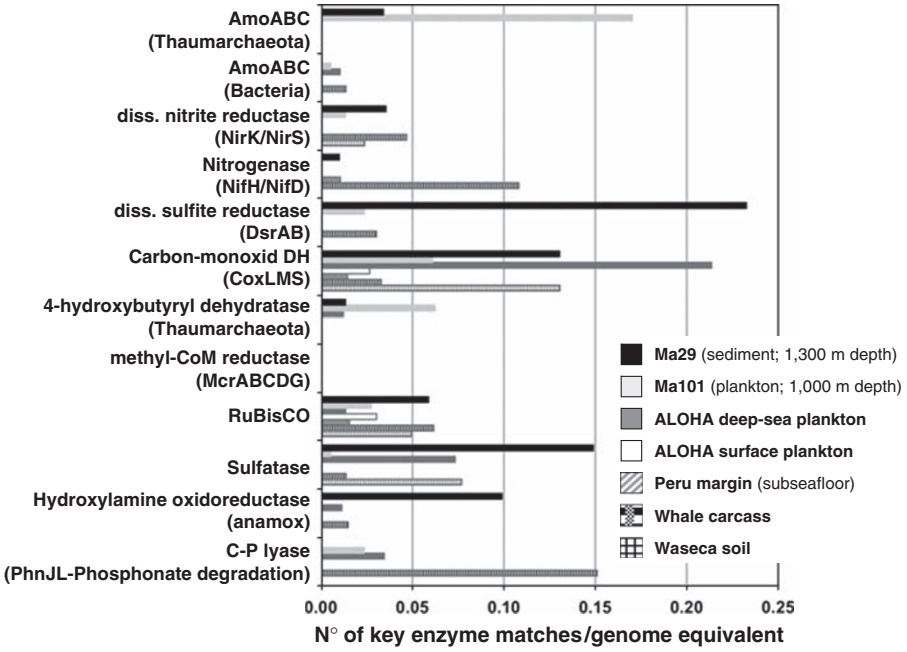

Figure 6 Comparison of the relative proportion of key metabolic enzymes in Marmara deep-sea plankton and sediment with other metagenomes.

plankton and soil, but not in the ALOHA water column (Table S2). The absence of genes for nitrate respiration in the highly oxygenated ALOHA water column in the open Pacific Ocean correlates with the fact that denitrification occurs when oxygen is limited and nitrate is available as terminal electron acceptor. This is the situation in the above suboxic to anoxic samples included in our study (deep Marmara Sea, whale carcass, soil). Furthermore, in the case of the whale carcass, the presence of these genes correlates with the occurrence of a significant number of SSU rRNA hits to Epsilonproteobacteria affiliated to *S. denitrificans* (22% of the whale carcass metagenome), which is able to perform denitrification. Tringe *et al*. (2005) already noted an enrichment of nitrate respiration genes in whale carcass and soil. Along with genes involved in nitrate respiration, genes involved in nitrogen fixation, a process inhibited by oxygen, were also relatively abundant in soil and Ma29 sediment metagenomes (Figure 6, Supplementary Table S2). Many of the nitrogen fixers in the whale carcass metagenome might be Alpha- or Epsilon-proteobacteria (Johnston *et al*., 2005).

As expected, genes encoding the subunits of the dissimilatory sulfite reductase responsible for the reduction of sulfite to sulfide during sulfate reduction were very abundant in the sediment Ma29, but were also present, though at much lower abundance, in Marmara deep-sea plankton and the whale carcass metagenomes (Figure 6 and Supplementary Table S2). Sulfate reduction is predominantly carried out by Deltaproteobacteria, which indeed accounted for ca 20% SSU rRNA genes in libraries and a similar percentage of rRNA hits in pyrosequences (Figure 1). Phylogenetic analyses of SSU rDNA sequences from gene libraries showed that many of the Ma29 Delta-proteobacteria affiliated to the sulfate-reducing Desulfobacteraceae, but also to several lineages without cultivated members (Supplementary Figure S7), which suggest that some of them may be sulfate reducers as well. Among the planktonic deltaproteobacterial SSU rDNA sequences, several belonged to the uncultivated group SAR324. The co-occurrence of genes for sulfate reduction in the same sample might suggest that the SAR324 are capable of reducing sulfate. Indeed, the presence of certain metabolic genes in metagenomic clones (Moreira *et al*., 2006) and the relative abundance of this group in oxygen-depleted waters (Zaikova *et al*., 2010) suggested that SAR324 may correspond to anaerobic or microaerophilic organisms.

In anoxic sediments, sulfate reduction is generally accompanied by the activity of methanogenic archaea in deeper sediment layers. In cold seep environments, such as those existing in localized areas of the Marmara Sea, some sulfate-reducing bacteria are symbiotically associated to archaea carrying out anaerobic methane oxidation. We thus looked for the presence of methyl coenzyme M reductase, which catalyses the terminal step in methanogenesis and seems to have a role also in reverse methanogenesis, being characteristic of methane-metabolizing archaea. We used the genes encoding the subunits of methyl coenzyme M from *Methanosarcina barkeri* (McrABCDG) and those of the nickel protein involved in the anaerobic oxidation of methane (McrABG) from an uncultured archaeon (Kruger *et al*., 2003) as seeds against the chosen metagenomes. However, no hits were detected. This is in agreement with both, the fact that Ma29 corresponded to 'normal', bottom sediment not strongly influenced by cold seep activity and with the fact that it was collected from the surface of the sediment core and hence above the methanogenesis layer. Accordingly, SSU rRNA sequences belonging to either typical methanogenic archaea or to ANME groups were scarce (<10%; Figure 1).

As Planctomycetes were relatively abundant in the Marmara sediment, we searched for the presence of genes that could indicate the occurrence of anammox activity by blasting 28 genes closely related to the *Kuenenia stuttgartensis* gene encoding the hydroxylamine oxidoreductase, one of the key enzymes of the anammox reaction (Strous *et al*., 2006). The highest number of matches was found in the Ma29 metagenome (8) followed by the whale carcass (2) and the ALOHA deep-sea metagenome (2). Although the number of matches remains relatively low, their relative higher abundance in Marmara sediment together with the presence of a relative high abundance of planctomycetes SSU rDNAs suggest that at least a fraction of these bacteria are able to oxidize ammonia anaerobically. Correlating with a relative high abundance of Planctomycetes in Marmara sediment, we also detected a high number of sulfatases (Figure 6 and Supplementary Table S2). Sulfatases are abundant in the genome of *R. baltica* (109 hits; Glockner *et al*., 2003) and, in general, marine planctomycetes possess a large number of these enzymes, which they might use for the initial breakdown of sulfated heteropolysaccharides, thus having an important role in recycling these abundant oceanic compounds (Woebken *et al*., 2007).

The use of phosphonate compounds has been recently proposed as an important source of phosphorous in P-depleted surface marine waters, as well as in more P-rich deep waters. Known genes but also novel pathways for phosphonate utilization are abundant in metagenomic picoplankton libraries as deduced from functional screenings (Martinez *et al*., 2009). Genes for phosphonate utilization were also abundant in the highly oligotrophic surface waters of the East Mediterranean (Feingersch *et al*., 2009). We looked for the presence of C-P lyase involved in phosphonate metabolism in our selected metagenomes. We detected a similar moderate abundance of homologs in Marmara and ALOHA deep-sea plankton. However, they were much more abundant in the whale carcass (Figure 6), which is

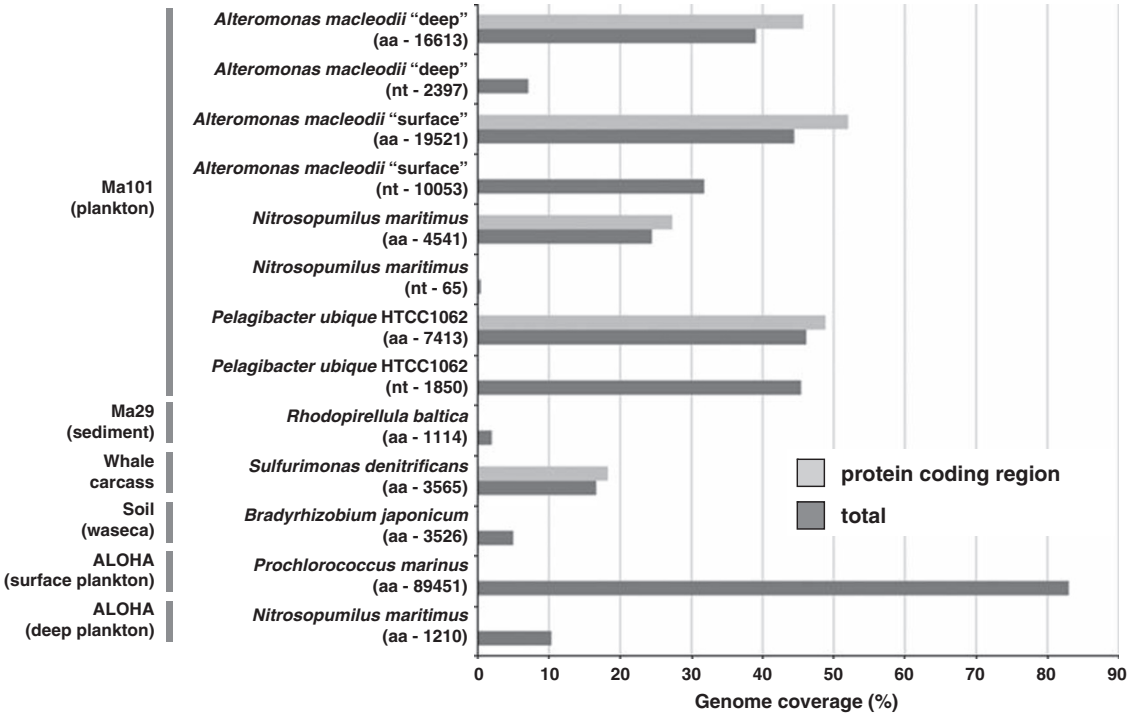

Figure 7 Percentage of coverage of the most abundantly recruited genomes in Marmara deep-sea plankton (Ma101) and sediment (Ma29) as compared with those most represented in other selected metagenomes. The numbers in parentheses indicate the numbers of matching reads; aa, based on amino acids (promer); nt based on nucleotides (nucmer).

most likely related to the active mobilization by these microbial communities of bone-associated phosphonates.

Finally, we also looked for carbon monoxide dehydrogenase genes. They were highly abundant in the ALOHA deep-sea metagenome (34 hits) followed by the Marmara Ma101 plankton and Ma29 sediment (19 hits each), soil (17 hits), the ALOHA surface metagenome (15 hits), the whale carcass (8) and Peru subsurface (4) metagenomes. The oxidation of CO to CO_2 as an alternative or supplementary energy source is widespread in many marine bacteria including, notably, members of the highly versatile and abundant *Roseobacter* clade (King and Weber, 2007; Brinkhoff *et al.*, 2008). Carbon monoxide dehydrogenase genes were detected at relative high abundance in deep Mediterranean waters, which suggested that deep-sea microbes might perform a similar form of lithoheterotrophy to that showed by surface bacterioplankton by oxidizing CO (Martin-Cuadrado *et al.*, 2007). The possible role of CO oxidation in deep waters has been criticized because the source of CO would be unclear at that depth and because carbon monoxide dehydrogenase is also involved in some pathways of central C metabolism, for instance in acetogenic methanogens (Pezacka and Wood, 1984). However, hydrothermal activity associated with oceanic ridges and to submarine volcanic areas, which are indeed rather extensive in the Mediterranean, constitutes a very likely source of CO in the deep sea. Furthermore, sequencing of metagenomic fosmids containing carbon monoxide dehydrogenase genes has shown that they are organized in clusters having the typical structure of CO oxidizing bacteria (Martin-Cuadrado *et al.*, 2009). Hence, lithoheterotrophy based on CO oxidation may actually be an important strategy to gain free energy also in the deep sea.

Genome recruitment in Marmara deep-sea plankton and sediment metagenomes

To get a deeper insight on the genomic structure of dominant microorganisms in Marmara Sea, we performed recruitment analysis using the promer and nucmer packages included in MUMmer 3.0 (Kurtz *et al.*, 2004). We selected as seeds the microbial genomes that showed the highest number of hits in the Marmara metagenomes. In Ma101 plankton, the microorganisms that retrieved more hits in pyrosequences with high average amino-acid identities were *Nitrosopumilus maritimus, P. ubique, A. macleodii* 'deep ecotype' and *A. macleodii* ATCC27126, isolated from surface waters (with 77%, 81%, 83% and 90% amino acid identities, respectively). Thus, we plotted these genome sequences against the Ma101 deep-sea plankton metagenome. At the amino-acid level (promer), the genomes of *A. macleodii* 'deep ecotype', *A. macleodii* ATCC27126, *P. ubique* and *N. maritimus* recruited Ma101 pyrosequences covering 39%, 44%, 46% and 25%, respectively, of the respective genomes and, when intergenic regions and ribosomal RNA genes

were excluded, the coverage increased to 45.7%, 52.0%, 48.8% and 27.3%, respectively (Figure 7). Similarly, at the nucleotide level (nucmer), we observed coverages of 31.8% (*A. macleodii* ATCC27126, 'surface ecotype'), 7.1% (*A. macleodii* 'deep ecotype'), 21.7% (*P. ubique*) and 0.435% (*N. maritimus*). The regions from *A. macleodii* 'deep-ecotype' not covered by Ma101 corresponded essentially to the (meta)genomic islands containing variable regions that were previously identified as major differences between this strain and the ATCC27126 (Ivars-Martinez *et al.*, 2008). Similarly, several (meta)genomic islands were identified in the recruitment plot of *P. ubique* (Supplementary Figure S22).

The dominance of *A. macleodii* ATCC27126, the so-called 'surface ecotype', over the 'deep-ecotype' in deep Marmara waters is at odds with some previous observations in other Mediterranean waters. Whereas the surface ecotype is dominant in surface Mediterranean and also in open ocean waters and seems to prefer warmer temperatures, the deep-ecotype is also present in surface waters, but in lower amounts and exhibits a preference for lower temperatures (Ivars-Martinez *et al.*, 2008). The 'deep-ecotype' isolate was obtained from 1000 m-deep Adriatic water at an average temperature of 13 °C, exactly the same depth and temperature (14 °C) as that of the Marmara Ma101 sample studied here. However, in Marmara, the surface ecotype predominates. It has been proposed that the 'deep-ecotype' is better adapted to microaerophilic conditions and that it might have a better tolerance to heavy metals (Ivars-Martinez *et al.*, 2008). However, this is also at odds with our observations, as the deep Marmara waters are suboxic and the Marmara Sea is heavily polluted (Ünlüata *et al.*, 1990; Beşiktepe *et al.*, 1994; Çagatay *et al.*, 2009). Two, nonmutually exclusive, explanations could account for the dominance of the surface ecotype in the deep Marmara Sea. First, *A. macleodii* cells from surface East Mediterranean (Aegean) waters are transported to depth after the water mass entering Marmara through the Dardanelles. Therefore, their presence would attest to the Aegean nature of the deep Marmara waters. However, this would not necessarily explain its relative high abundance at this depth. The second explanation would refer to the higher metabolic flexibility of the surface ecotype, which is able to exploit a wider range of substrates (Ivars-Martinez *et al.*, 2008) and might eventually be more adapted to Marmara bathypelagic waters.

In Marmara sediment, recruitment was much poorer than in deep-sea plankton because of the larger microbial diversity and, hence, lower genome coverage. The genomes that recruited best were those of the planctomycetes *Pirellula* sp. 1 (2520 hits) and *Blastopirellula maris* (2133 hits), followed by that of *Geobacter uranioreducens* Rf4 (1853 hits), *Pelobacter carbinolicus* DSM2380 (1494 hits) and *Nitrosococcus oceani* (1375 hits), illustrating the high relative abundances of, mostly, planctomycetes and sulfate-reducing Deltaproteobacteria.

Concluding remarks

This study is a contribution to incipient comparative metagenomics of deep-sea microbial communities from both plankton and sediment to start building a better picture of the microbial diversity and the associated gene content and functional potential as a function of geographical and physico-chemical environmental parameters. We analyzed metagenomic 454 pyrosequences from bathypelagic plankton (Ma101, 1000 m depth) and from 'normal', non-cold-seep influenced, bottom sediment (Ma29, 1300 m bottom depth) in the Sea of Marmara. Overall, the metagenomic data were in good agreement with the bacterial diversity inferred from SSU rRNA gene libraries. Bathypelagic Marmara plankton was dominated by Gamma- and Alpha-proteobacteria with the 'surface' *A. macleodii* ecotype and *P. ubique* as dominant lineages, as shown by genome recruitment analyses. Within the archaea, group I Crenarchaeota/Thaumarchaeota that seem capable of oxidizing ammonia, as attested by the presence of archaeal *amo* genes, dominate. Sediment communities are extremely diverse, with particularly high relative abundances of Deltaproteobacteria, many of them are sulfate reducers, and planctomycetes, which are important in carbon cycling. Methanogenic and/or methanotrophic archaea were not abundant in the surface sediment sample, indicating that the transition to the methanogenesis zone is deeper and that the sediment is not actually influenced by intense seeping activity, which, in Marmara, is restricted to localized black patches along the main North Anatolian Fault (Géli *et al.*, 2008; Zitter *et al.*, 2008). Future comparison with metagenome sequences from those seep areas in the same location would be interesting to reveal specific adaptations to those environments. Although a large diversity of eukaryotes was detected from both plankton and sediment, eukaryotic sequences (including rRNA genes) in the metagenome data sets were much less abundant than prokaryotic ones. The comparison of the Marmara metagenomic data with those of other deep-sea data sets (ALOHA deep-sea plankton, whale carcasses), marine subsurface sediment of the Peru margin, and other outgroup environments (ALOHA surface plankton and soil) showed that Marmara deep-sea plankton resembled both the meso- and bathypelagic Pacific ALOHA deep-sea and the whale carcass metagenomes. Although Marmara bathypelagic waters share the deep-sea conditions with the ALOHA deep-sea plankton, it shares suboxic conditions with the chemosynthetic ecosystem of whale carcasses. Marmara sediment clusters with the soil metagenome, indicating that the ecological role involving the ultimate degradation of organic matter and the completion of

biogeochemical cycles imposes strong constraints and determines the nature and function of the associated microbial communities.

Acknowledgements

We thank P Henry and N Çagatay, chief scientists of the oceanographic cruise *MARNAUT*, for providing PLG the opportunity to participate, as well as to the shipboard scientific team for an enjoyable cruise. We also acknowledge Captain and crew of *R/V L'Atalante* and the help of the Turkish Navy to protect our ship in the zones of heavy ship traffic. We thank INSU for providing a CTD-rosette for water sampling and L Fichen, who operated it. This work was funded by the French National Agency for Research (EVOLDEEP project, contract number ANR-08-GENM-024-002).

References

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

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.

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.

Beşiktepe ŞT, Sur İH, Özsoy E, Abdul Latif M, Oğuz T, Ünlüata Ü. (1994). The circulation and hydrography of the Marmara Sea. *Prog Oceanogr* **34**: 285–334.

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH. (2008). Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* **105**: 10583–11058.

Blackburn MV, Hannah F, Rogerson A. (2009). First account of apochlorotic diatoms from intertidal sand of a south Florida beach. *Estuarine Coastal and Shelf Science* **84**: 519–526.

Brinkhoff T, Giebel HA, Simon M. (2008). Diversity, ecology, and genomics of the *Roseobacter* clade: a short overview. *Arch Microbiol* **189**: 531–539.

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

Çagatay MN, Eris K, Ryan WBF, Sancar Ü, Polonia A, Akçer S *et al.* (2009). Late Pleistocene-Holocene evolution of the northern shelf of the Sea of Marmara. *Mar Geol* **265**: 87–100.

Cetecioglu Z, Ince BK, Kolukirik M, Ince O. (2009). Biogeographical distribution and diversity of bacterial and archaeal communities within highly polluted anoxic marine sediments from the marmara sea. *Mar Pollut Bull* **58**: 384–395.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

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.

DeSantis Jr TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM *et al.* (2006b). NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res* **34**: W394–W399.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006a). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Feingersch R, Suzuki MT, Shmoish M, Sharon I, Sabehi G, Partensky F *et al.* (2009). Microbial community genomics in eastern Mediterranean Sea surface waters. *ISME J* **4**: 78–87.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105**: 3805–3810.

Frigaard NU, Martinez A, Mincer TJ, DeLong EF. (2006). Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847–850.

Géli L, Henry P, Zitter T, Dupré S, Tryon M, Çağatay MN *et al.* (2008). Gas emissions and active tectonics within the submerged section of the North Anatolian Fault zone in the Sea of Marmara. *Earth and Planetary Science Letters* **274**: 34–39.

Glockner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W *et al.* (2003). Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci USA* **30**: 30.

Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Ivars-Martinez E, Martin-Cuadrado AB, D'Auria G, Mira A, Ferriera S, Johnson J *et al.* (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J* **2**: 1194–1212.

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

Johnston AW, Li Y, Ogilvie L. (2005). Metagenomic marine nitrogen fixation—feast or famine? *Trends Microbiol* **13**: 416–420.

Jorgensen BB, Boetius A. (2007). Feast and famine—microbial life in the deep-sea bed. *Nat Rev Microbiol* **5**: 770–781.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–D280.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.

King GM, Weber CF. (2007). Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat Rev Microbiol* **5**: 107–118.

Konneke M, Bernhard AE, de la JR , Walker CB, Waterbury JB, Stahl DA. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.

Konstantinidis KT, Braff J, Karl DM, DeLong EF. (2009). Comparative metagenomic analysis of a microbial community residing at a depth of 4000 meters at

station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* **75**: 5345–5355.

Kruger M, Meyerdierks A, Glockner FO, Amann R, Widdel F, Kube M *et al.* (2003). A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. *Nature* **426**: 878–881.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW *et al.* (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806–809.

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

Lovejoy C, Massana R, Pedros-Alio C. (2006). Diversity and distribution of marine microbial eukaryotes in the Arctic Ocean and adjacent seas. *Appl Environ Microbiol* **72**: 3085–3095.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Martin-Cuadrado AB, Ghai R, Gonzaga A, Rodriguez-Valera F. (2009). CO dehydrogenase genes found in metagenomic fosmid clones from the deep Mediterranean. *Appl Environ Microbiol* **75**: 7436–7444.

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.

Martin-Cuadrado AB, Rodriguez-Valera F, Moreira D, Alba JC, Ivars-Martinez E, Henn MR *et al.* (2008). 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. *ISME J* **2**: 865–886.

Martinez A, Tyson GW, Delong EF. (2009). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* **12**: 222–238.

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.

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.

Pezacka E, Wood HG. (1984). Role of carbon monoxide dehydrogenase in the autotrophic pathway used by acetogenic bacteria. *Proc Natl Acad Sci USA* **81**: 6261–6265.

Polymenakou PN, Bertilsson S, Tselepides A, Stephanou EG. (2005). Bacterial community composition in different sediments from the Eastern Mediterranean Sea: a comparison of four 16S ribosomal DNA clone libraries. *Microb Ecol* **50**: 447–462.

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11**: 1358–1375.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Quaiser A, Lopez-Garcia P, Zivanovic Y, Henn MR, Rodriguez-Valera F, Moreira D. (2008). Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton. *Environ Microbiol* **10**: 2704–2717.

Raes J, Foerstner KU, Bork P. (2007). Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10**: 490–498.

Ruehland C, Blazejak A, Lott C, Loy A, Erseus C, Dubilier N. (2008). Multiple bacterial symbionts in two species of co-occurring gutless oligochaete worms from Mediterranean sea grass sediments. *Environ Microbiol* **10**: 3404–3416.

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.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N *et al.* (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**: 374–378.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Schmidt TM, DeLong EF, Pace NR. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* **173**: 4371–4378.

Shi Y, Tyson GW, DeLong EF. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**: 266–269.

Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW *et al.* (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790–794.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

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.

Ünlüata Ü, Oðuz T, Latif MA, Özsoy E. (1990). On the physical oceanography of the Turkish Straits. In: Pratt LJ (ed). *The Physical Oceanography of Sea Straits*. Kluver: Dordrecht. pp 25–60.

Urich T, Lanzen A, Qi J, Huson DH, Schleper C, Schuster SC. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**: e2527.

Varela MM, van Aken HM, Sintes E, Herndl GJ. (2008). Latitudinal trends of Crenarchaeota and Bacteria in the meso- and bathypelagic water masses of the Eastern North Atlantic. *Environ Microbiol* **10**: 110–124.

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.

Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Appl Environ Microbiol* **64**: 3075–3078.

Vetriani C, Jannasch HW, MacGregor BJ, Stahl DA, Reysenbach AL. (1999). Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl Environ Microbiol* **65**: 4375–4384.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ *et al.* (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

Wagner-Dobler I, Biebl H. (2006). Environmental biology of the marine *Roseobacter* lineage. *Annu Rev Microbiol* **60**: 255–280.

Woebken D, Teeling H, Wecker P, Dumitriu A, Kostadinov I, Delong EF *et al.* (2007). Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J* **1**: 419–435.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Wright TD, Vergin KL, Boyd PW, Giovannoni SJ. (1997). A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Appl Environ Microbiol* **63**: 1441–1448.

Yakimov MM, La Cono V, Denaro R, D'Auria G, Decembrini F, Timmis KN *et al.* (2007). Primary producing prokaryotic communities of brine, interface and seawater above the halocline of deep anoxic lake L'Atalante, Eastern Mediterranean Sea. *ISME J* **1**: 743–755.

Zaikova E, Walsh DA, Stilwell CP, Mohn WW, Tortell PD, Hallam SJ. (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ Microbiol* **12**: 172–191.

Zitter TAC, Henry P, Aloisi G, Delaygue G, Çagatay MN, Mercier de Lepinay B *et al.* (2008). Cold seeps along the main Marmara Fault in the Sea of Marmara (Turkey). *Deep-Sea Research I* **55**: 552–570.

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