The ISME Journal (2010) 4, 1154–1166
© 2010 International Society for Microbial Ecology All rights reserved 1751-7362/10
www.nature.com/ismej

ORIGINAL ARTICLE

Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing

Rohit Ghai[1], Ana-Belén Martin-Cuadrado[1], Aitor Gonzaga Molto[1], Inmaculada García Heredia[1], Raúl Cabrera[2], Javier Martin[3], Miguel Verdú[3], Philippe Deschamps[4], David Moreira[4], Purificación López-García[4], Alex Mira[2] and Francisco Rodriguez-Valera[1]

[1]*Departmento de Producción Vegetal y Microbiología, Evolutionary Genomics Group, Universidad Miguel Hernández, Apartado 18 San Juan de Alicante, Alicante, Spain;* [2]*Department of Genomics and Health, Centro Superior de Investigación en Salud Pública, Avda. Cataluña 21, Valencia, Spain;* [3]*Mediterraneo Servicios Marinos S.L. Nueva Dársena Pesquera s/n, Alicante, Spain;* [4]*Unité d'Ecologie, Systématique et Evolution, CNRS UMR8079, Université Paris-Sud 11, Orsay, France*

The deep chlorophyll maximum (DCM) is a zone of maximal photosynthetic activity, generally located toward the base of the photic zone in lakes and oceans. In the tropical waters, this is a permanent feature, but in the Mediterranean and other temperate waters, the DCM is a seasonal phenomenon. The metagenome from a single sample of a mature Mediterranean DCM community has been 454 pyrosequenced both directly and after cloning in fosmids. This study is the first to be carried out at this sequencing depth (ca. 600 Mb combining direct and fosmid sequencing) at any DCM. Our results indicate a microbial community massively dominated by the high-light-adapted *Prochlorococcus marinus* subsp. *pastoris*, *Synechococcus* sp., and the heterotroph *Candidatus* Pelagibacter. The sequences retrieved were remarkably similar to the existing genome of *P. marinus* subsp. *pastoris* with a nucleotide identity over 98%. Besides, we found a large number of cyanophages that could prey on this microbe, although sequence conservation was much lower. The high abundance of phage sequences in the cellular size fraction indicated a remarkably high proportion of cells suffering phage lytic attack. In addition, several fosmids clearly belonging to Group II Euryarchaeota were retrieved and recruited many fragments from the total direct DNA sequencing suggesting that this group might be quite abundant in this habitat. The comparison between the direct and fosmids sequencing revealed a bias in the fosmid libraries against low-GC DNA and specifically against the two most dominant members of the community, *Candidatus* Pelagibacter and *P. marinus* subsp. *pastoris*, thus unexpectedly providing a feasible method to obtain large genomic fragments from other less prevalent members of this community.
The ISME Journal (2010) **4,** 1154–1166; doi:10.1038/ismej.2010.44; published online 15 April 2010
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metagenomics; pyrosequencing; *Prochlorococcus*; Pelagibacter; cyanophages; deep chlorophyll maximum

Introduction

The sequencing of metagenomic DNA has followed until recently two main routes derived from the use of different cloning vectors. The cloning of the DNA in small insert vectors (plasmids), often known as shotgun cloning, has allowed the sequencing of large libraries and the accumulation of vast amounts of sequence information about some samples such as the Sargasso Sea (Venter *et al.*, 2004). As an alternative, some studies have been carried out cloning the environmental DNA into large insert vectors (mostly fosmids) and end sequencing the libraries through vector primers (Beja *et al.*, 2000b; DeLong *et al.*, 2006; Martin-Cuadrado *et al.*, 2007). This provides lower taxonomic coverage of the samples, but allows to fully sequence fosmids that promise interesting insights in the community. The average fosmid size of ca. 35 kb provides information about natural gene clusters (for example bacterial operons) and allows to infer functionality much more precisely than the individual genes retrieved with short insert vectors (Martin-Cuadrado *et al.*, 2009). Besides, the phylogenetic position of the microbe represented by the fosmid can be

Correspondence: F Rodriguez-Valera, Departamento Producción Vegetal y Microbiología, Evolutionary Genomics Group, Universidad Miguel Hernández, San Juan de Alicante, Alicante 03350, Spain.
E-mail: frvalera@umh.es
Received 30 November 2009; revised 2 March 2010; accepted 15 March 2010; published online 15 April 2010

ascertained much more reliably, either from the presence of housekeeping genes, or when a consensus of similarity is found over many of the genes to some specific phylogenetic group or by the oligonucleotide usage of the fosmid with that of fully sequenced genomes.

The 454 pyrosequencing opens a third route, that of direct sequencing (DS) without cloning in a cellular host, instead the emulsion PCR amplification step is carried out on beads. The most recent platforms provide reads as long as 400 bp, which allows a much more reliable annotation of the fragments, and, depending on the diversity present in the sample, assembly of larger contigs compared with earlier 454 pyrosequencing standards. However, given that assembling large contigs from natural samples, especially of the less-abundant organisms, has proven to be difficult even when vast amounts of sequence were available (Venter *et al.*, 2004), the large insert vector approach has still some advantages, principally because these DNA fragments contain naturally linked genomic features (that is belonging to the same microbe). Large insert vectors can also be efficiently sequenced by 454 pyrosequencing by pooling clones and sequencing them as separate samples in a pyrosequencing plate. If the number of fosmids per pool is low, long fragments could theoretically be assembled. In an attempt to compare the efficiency of both approaches to describe a complex natural habitat, we have applied both methodologies to the same water sample from the Mediterranean deep chlorophyll maximum (DCM).

A DCM corresponds to the layer of maximal chlorophyll concentration in the water column of oceans or lakes. The simultaneous occurrence of relatively high inorganic nutrient concentration and appropriate wavelength and intensity of light here provides ideal conditions for phytoplankton development, and it is actually at this section in the water column where most of the photosynthetic primary productivity takes place and the highest population density of marine microbes is found (Estrada *et al.*, 1993). Throughout much of the tropical ocean, the DCM is a permanent feature. However, at higher latitudes, it occurs seasonally as winter mixing of the upper 100–200 m prevents stratification. Over most of the Mediterranean, the DCM typically forms during April–May and is maintained as a very sharp feature of the water column till late October–November. We have extracted metagenomic DNA from the picoplanktonic 5–0.2 μm size fraction (mostly prokaryotic cells) from a single sample obtained in mid-October at 50 m depth and analyzed the DNA from this sample (without cloning or amplification) by 454 pyrosequencing (hereafter referred to as DS). We also used 454 to sequence ca. 1152 randomly selected fosmid clones from a metagenomic fosmid library constructed from the same DNA sample. Here, we compare the output from both methods and the picture of the microbial community that they provide.

Materials and methods

Sample collection and processing

A single seawater was collected on 15 October 2007 from the DCM layer (50 m deep) off the coast of Alicante, Spain (38° 4′6.64″ N/0° 13′55.18″ W) with a Niskin bottle. Some physico-chemical parameters are shown in Supplementary Table T6. The sample was sequentially filtered through a 5 μm pore size polycarbonate filter and 0.22 μm pore size Sterivex filters (Durapore, Millipore, Billerica, MA, USA) using a peristaltic pump. Sterivex filters (retaining the 0.2–5 μm diameter planktonic cells) were filled up with lysis buffer (40 mM EDTA, 50 mM Tris/HCl, 0.75 M sucrose) and conserved at −20 °C until DNA extraction. The lysed cells from sterivex filters collecting the biomass from ca. 30 l of sample were pooled and then DNA was extracted as described before (Martin-Cuadrado *et al.*, 2007). Briefly, filters were thawed on ice and then treated with 1 mg ml^{-1} lysozyme and 0.2 mg ml^{-1} proteinase K (final concentrations). Nucleic acids were extracted with phenol/chloroform/isoamyl alcohol and chloroform/isoamyl alcohol and DNA integrity was checked by agarose gel electrophoresis.

Fosmid library construction and sequencing strategy

The fosmid genomic library was constructed using the CopyControl Fosmid Library Production kit (Epicentre, Madison, WI, USA), as described by the manufacturer's instructions, from ~9.7 μg of DNA. The fosmid library generated had a total of 12 192 clones. From this collection, 1152 (12 96-well plates) were randomly selected for 454 pyrosequencing. Clones were individually grown and induced using 1 ml 96-well plates. All wells within each plate were pooled before DNA extraction. DNA was extracted using the QIAprep Spin Miniprep kit (Qiagen, Valencia, CA, USA). A total of 5 μg of DNA from each pooled plate was sent separately for sequencing (Roche 454 GS-FLX system, Titanium chemistry, by GATC, Konstanz, Germany). DNA from each plate was tagged individually using a multiplex identifier adaptor containing a unique 10-base pair sequence that is recognized by the sequencing analysis software, allowing for automated sorting of multiplex identifier adaptor-containing reads, and facilitating the assembly. In parallel, 5 μg from the same DNA batch used for fosmid cloning were sent for direct pyrosequencing (DS).

Annotation and assembly

Assembly was performed using the CLC Genomics Workbench v. 3.5 with minimum %identity of 95% and a sequence overlap of at least 25% of the read length. Gene prediction was performed using MGA (Noguchi *et al.*, 2008). The predicted protein sequences obtained were compared using blastp to the NCBI-NR protein database. The DCM DS data set was also annotated using the MG-RAST server (Meyer *et al.*, 2008).

Community structure using all reads
For taxonomy, the data sets were compared using BLASTN (Altschul *et al*., 1997) to a combined database containing the NCBI-NT database and whole genome shotgun assembly data for 1000 draft microbial genomes from NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The data were analyzed using MEGAN (Huson *et al*., 2007).

Community structure using ribosomal RNA
Ribosomal RNA (rRNA) genes were identified by comparing the data sets against the RDP (Cole *et al*., 2009) and the SILVA long ribosomal subunit database (Pruesse *et al*., 2007). All reads that matched an rRNA sequence with an identity $>90\%$ and an alignment length of >200 bases against either the RDP or the LSU database were extracted. The best hit with a taxonomic affiliation was considered a reasonable closest attempt to classify the rRNA sequences.

Recruitment plots
Fragment recruitment of the DCM DS data set was performed against all complete and draft microbial genomes using BLASTN. However, all rRNA sequences (16S, 23S, and intergenic spacers) were removed from the DCM DS data sets before performing any recruitment. The criteria for counting a hit were minimum %identity of 95% and minimum alignment of 50 bp. The same criteria were used when performing recruitment of the DCM DS data by the assembled DCM fosmids and against all known marine viral genomes. Data were plotted using R (http://cran.r-project.org).

Comparison with other metagenomic data sets
The DCM DS data was compared against the entire global ocean sampling (GOS) expedition data (Rusch *et al*., 2007) using BLASTN, and a hit was counted using the criteria of minimum 95% identity and alignment length of 50 bases. For comparison with SEED subsystems (http://www.theseed.org), the DCM DS data set was annotated using the MG-RAST server (Meyer *et al*., 2008). The already available annotation for the GOS data sets was used from the same server to gather their SEED subsystem classification. The relative abundance of different subsystems were calculated as a $\log_2$ ratio as described before (Konstantinidis *et al*., 2009).

Phylogenetic analysis of archaeal fosmids and rhodopsin sequences
From the first annotation of the fosmids, using best BLAST hit analysis, we identified a set of prospective archaeal fosmids. To confirm their phylogenetic affiliation, we ran a BLASTp for every predicted ORF encoded in each fosmid against a protein database containing 394 full genome sequences (291 bacteria, 50 archaea, 53 eukaryotes). From each BLAST result, when possible, we retrieved the identity of the best hit and up to 50 protein sequences of the top hits with an e-value $\leqslant$1e-05. We then constructed an alignment of these protein sets using the program Muscle (Edgar, 2004) and reconstructed the corresponding phylogenetic trees by Maximum Likelihood using the program Treefinder (Jobb *et al*., 2004), with the substitution model WAG and a Γ law to deal with unequal evolutionary rates among sites. Bootstraps values were obtained from 1000 replicates. Every tree was manually analyzed to determine the phylogenetic relationship of the sample sequence. To look for rhodopsin genes, we used several representative rhodopsin protein sequences to search in the DCM DS. A total of 39 reads were found to match our rhodopsin sequences at a similarity level of $>70\%$ and alignment length >100 amino acids. We aligned our sequences with the closest sequences in the database at the amino-acid level and then constructed a maximum likelihood phylogenetic tree as mentioned above using 77 non-ambiguously aligned positions.

Nucleotide accession numbers. The assembled fosmid sequences have been deposited in Genbank (GU942957–GU943153). The raw data from the DS data set are available from the NCBI Short Read Archive (SRP002017).

Results

General features of the sequence generated by both approaches
Both 454 pyrosequencing plates provided close to 1 million reads. The average length of the reads was shorter in the DS data set (260 bp) than in the pooled fosmids plate (358 bp). However, nearly one-fourth of the sequence data obtained from the fosmids corresponded to the vector, thus the amount of environmental sequence data from both data sets was similar (ca. 312 Mbp for DS and 325 Mbp for the fosmids). Figure 1 shows the GC% distribution plot of the individual reads in the data sets together with a GOS data set that included sequences from a similar filter size range (0.1–3 μm) to ours (GS001b and GS001c). The DS data set clearly shows a peak at ca. 34% GC and is very similar to the GOS sample distribution. On the other hand, the DCM fosmid reads had two peaks of similar height at 35% and 52%. Actually in both the DS and the GOS plots there is a slight shoulder at the value of the fosmids high-GC peak, so DNA of this GC content is present in both data sets as well, but seems to be enriched in the fosmids. Similar two-peak GC plots have been described also for BAC libraries (Feingersch and Béjà, 2009) and might reflect a cloning bias.

To compare the DNA sequences retrieved in the DS and fosmid data sets, we performed comparisons

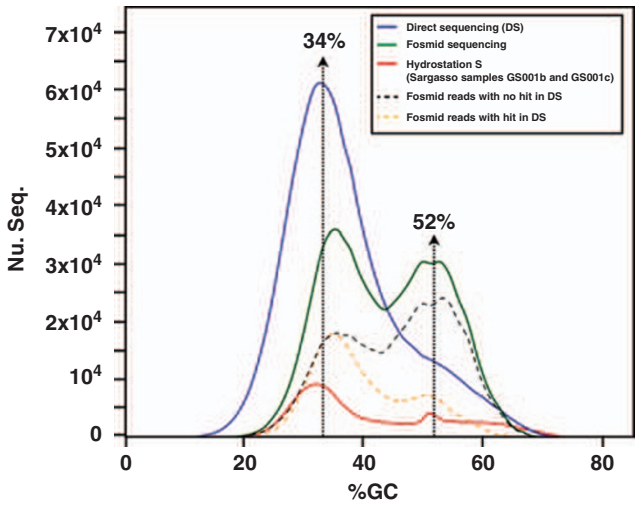


Figure 1 GC% distribution of the sequences generated. *Blue*: All reads obtained from direct sequencing (DS) data set; *green*: all reads obtained from sequencing of the fosmids, after removal of the vector sequence; *red*: reads from GOS sample Sargasso Sea Hydrostation S (0.1–0.8 and 0.8–3.0 μm combined); *orange dashes*: reads from fosmids that have at least one hit in the direct sequencing (DS) data set; *black dashes*: reads from fosmids that have no hit in the direct sequencing (DS) data set.

of the unassembled reads using BLAST, at high identity cutoff (>95%) and different minimum alignment lengths (25, 50, and 100 bp). The reads of DS having hits in the fosmids was only 2.7% at a minimum alignment length of 100 bp. This is expected as the fosmids likely represent a small fraction of the diversity present in the sample (assuming an average fosmid size of 30 kb, we can estimate that only about 35 Mb or ~10 genome equivalents of environmental DNA could be inserted in the 1152 fosmids sequenced) versus the likely hundreds of different genomes that could be present in the natural sample, and could be at least partially retrieved by DS. However, the relatively small number of reciprocal hits, that is fosmid reads with hits in the DS (20.2% with overlap of at least 100 nucleotides and only up to 33.1% decreasing the overlap to 25 nucleotides), came as a surprise. This might indicate that the 300 Mb retrieved by DS is still far from providing an adequate coverage of most of the microbes present. Besides, fosmid libraries could also capture genomic fragments from less predominant members of the community (see below). The GC plot of Figure 1 points in that direction as the majority of fosmid reads that do not have hits in the DS data set tend to be high GC and those that do have hits are primarily low GC.

Community structure

Individual reads. One of the important objectives in all metagenomic studies is to establish the microbial community structure in natural samples. Here, we have compared the snapshot community structure provided by each of the approaches and also derived a consensus from both. The rRNA genes are often used to get reliable phylogenetic affiliation because of the vast database of their sequences and relatively high conservation. In our case, the DS reads yielded 1058 sequences that could be assigned to either the 16S or the 23S rRNA genes. The picture obtained from this analysis shows the community to be massively dominated by the α-proteobacterium *Candidatus* Pelagibacter and the picocyanobacteria *Prochlorococcus marinus* and *Synechococcus* sp., followed in smaller amounts by uncultivated group II Euryarchaeota (Figure 2). Median %identity for rRNA reads identifying *P. marinus* and *Candidatus* Pelagibacter were 99% and 97%, respectively. The rRNA genes from Crenarchaeota were not detected. This description fits quite well with earlier analysis of the community structure of the photic zone obtained by metagenomics (DeLong *et al.*, 2006; Feingersch *et al.*, 2010). PCR amplification of the 16S rRNA gene followed by cloning is strongly biased against picocyanobacteria and archaea, so that these groups seemed undervalued by this approach (see for example Zaballos *et al.*, 2006). The total number of reads in the fosmids data set that matched an rRNA sequence was much smaller (only ~100 reads). Nearly half of all these belonged to *Mantoniella squamata* (24 reads) and *Micromonas* sp. (21 reads), both widespread eukaryotic green algae (Vaulot *et al.*, 2008). Among the bacterial rRNAs, we obtained nearly 40 reads from *Prochlorococcus* and *Synechococcus*, although we were able to assemble only 17 contigs longer than 3 kb belonging to these microbes from the fosmid data set (and only three longer than 10 kb). The only other cyanobacterium for which we retrieved any rRNA sequences was *Cyanobium*, a marine or fresh-water relative of *Synechococcus* (two sequences). No rRNA sequences were identified for *Candidatus* Pelagibacter. The remaining few hits were found to be most similar to bacteria belonging to δ/ε-proteobacteria subdivisions (for example *Geopsychrobacter*, *Geobacter*, *Pelobacter*). No archaeal rRNA was found in the fosmids.

Another approach to analyze the taxonomic distribution of the fragments in a metagenomic data set is using MEGAN (Huson *et al.*, 2007). With this, we could assign a significant number of reads from both data sets (25% of the reads from the DS and 10% from the fosmids) (Figure 2). The majority of the identifiable reads in both data sets were assigned to bacteria, although there were more eukaryotic reads in the fosmids data set. Nearly no reads were ascribed to Archaea, as expected as genomes of Group II Euryarchaeota are not yet available. Regarding bacteria, in the two data sets, Proteobacteria had the maximum number of reads, followed by cyanobacteria (Figure 2), similar to what we found by classifying the rRNA genes from the DS data set. However, at a finer resolution, the fosmids data set had a higher percentage of reads assigned to γ-proteobacteria (54%), followed by α-proteobacteria

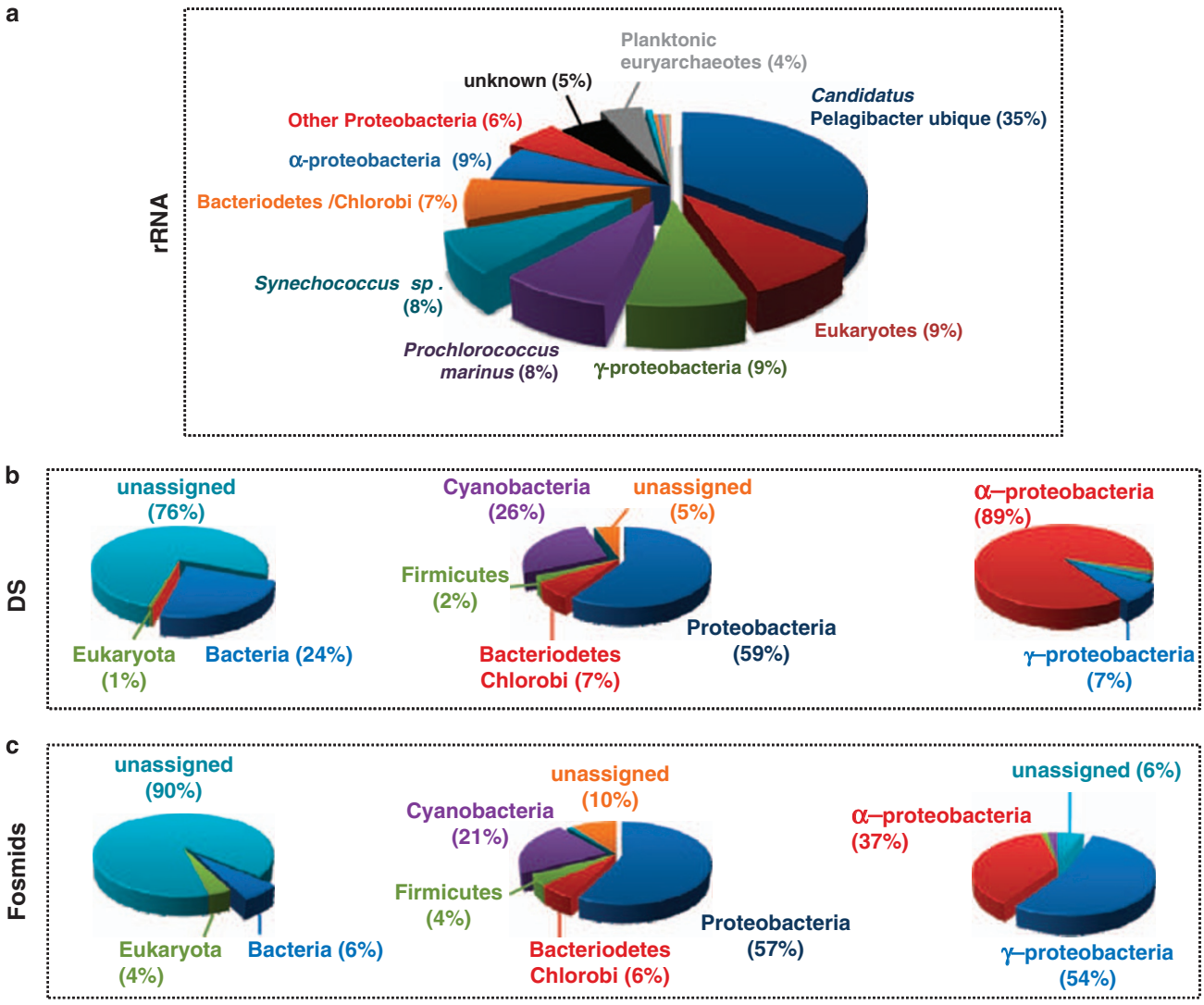


Figure 2 (**a**) Classification of the DCM DS reads using the RDP (16S) and the LSU databases (23S). (**b**, **c**) Comparison of the two data sets, direct pyrosequencing (DS), and fosmid sequencing, using BLASTN of the reads against the nucleotide database NT + 1000 draft microbial genomes, classified using MEGAN. *From left to right*: Kingdom level classification, classification of bacterial reads, and classification of proteobacterial reads.

(37%), whereas the DS reads showed the opposite (α-proteobacteria 89% and γ-proteobacteria 7%). Though this difference might in principle be due to insufficient sampling, it may also be an indication of a bias in the fosmid cloning against the α-proteobacterium *Candidatus* Pelagibacter, as has been suggested before by other authors (Feingersch and Béjà, 2009; Temperton *et al.*, 2009 and so on.). The δ- and β-proteobacteria were assigned only a few reads (~1% of the proteobacterial reads) in both data sets. The phylogenetic profile for the DS data set was also examined using the SEED subsystems database available through the MG-RAST server (Meyer *et al.*, 2008). The results obtained were similar (Supplementary Figure S1) to those described above. For a taxonomic overview of the sequences common to both data sets, see Supplementary Table T3.

Assembled sequences

The amount of sequence that could be assembled from both data sets is described in Supplementary Table 1. From the DS pyrosequencing only five contigs were larger than 3 kb (the largest being ~4.4 kb). Two of these clearly belonged to *P. marinus* subsp. *pastoris.* The similarity and synteny were so high that the fragments could be assigned at the strain level (>98% nucleotide identity). Other assembled contigs contained genes similar to cyanophage genes, specifically to the *Prochlorococcus* phage PSSM2 and the *Synechococcus* phages S-PM2 and syn9, albeit at much lower similarities. All are myoviruses with large genomes. PSSM2 was isolated from low-light-adapted *P. marinus* NATL1A, S-PM2 from *Synechococcus* WH7803, and syn9 from *Synechococcus* WH8012, but is also known to infect *Prochlorococcus* (Sullivan *et al.*, 2006).

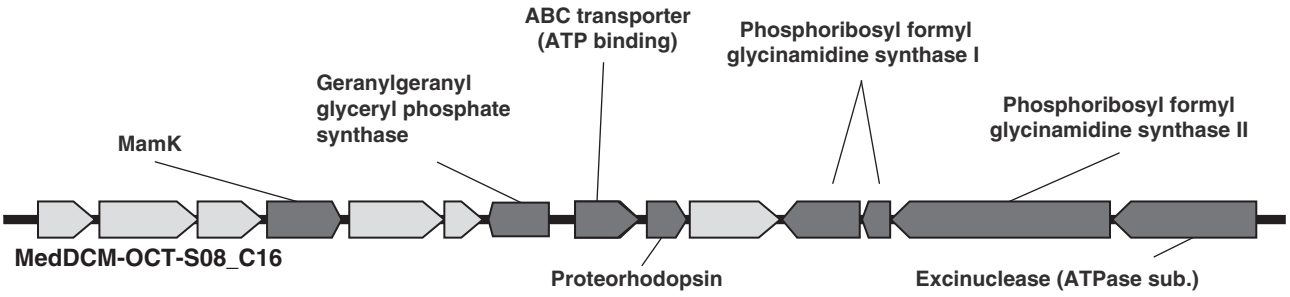
As expected, fosmids provided many large contigs with a total of 8.8 Mb assembled in 1287 contigs >3 kb (versus only ~18 kb assembled from the DS). However, only 55 contigs from the fosmid data set were over 20 kb, corresponding to the expected fosmid insert size. This indicates that the number of fosmids used was too large to get a proper and balanced coverage for most of them. Still, the large size of the fosmid contigs is advantageous, for example, to identify microbes, as a consensus can be derived from more than one gene. We could safely assign nearly all the assembled fosmid sequences >10 kb, as smaller sequences often provide contradictory or unreliable assignments. From now on we will limit the description to the 197 assembled fosmid sequences >10 kb.

The largest number of readily identifiable fosmids were from cyanophages (34) and the majority of the cyanophage fragments assembled from the fosmids belonged to the Cyanophage PSSP-7 (12) (family podoviridae) (which actually was isolated using the *P. marinus* subsp. *pastoris* strain MED4) (Lindell *et al.*, 2004). A number of fosmids (9) were also identified as belonging to Myoviruses (S-PM2, PSSM2). This indicates that cyanophages must constitute a significant fraction of the genetic material present in the DCM. The assembled cyanophage fosmids showed excellent synteny with existing reference genome of the cyanophage PSSP-7 (Supplementary Figure S2). But the similarity was again relatively low (60–70% nucleotide identity) compared with the extremely high similarity found for DS contigs and the MED4 genome. Only three fosmid contigs could be ascribed to cyanobacterial cellular DNA (two belonging to *Synechococcus* sp. CC9605 and one to *P. marinus* subsp. *pastoris*). Similar to the DS contigs, the fosmid belonging to *P. marinus* was over 97% nucleotide identity (similarity at the strain level). However, both the *Synechococcus* fosmids and the single *Candidatus* Pelagibacter fosmid were on the average 95% and 84% identical, respectively, that is belonged to microbes much more different to the sequenced strain of reference. Another large group of fosmids (44) were again phages, but could not be classified. They might be cyanophages belonging to distantly related groups or phages from other abundant members of the community.

Twenty-two fosmids could be clearly ascribed to marine Euryarchaeota. The abundance of this group in the Hawaii Ocean Time-Series (HOT) permanent DCM has been described before (DeLong *et al.*, 2006) by 16S rRNA gene amplification from a metagenomic fosmid library. Given the scarcity of sequence information about this group, the archaeal ascription of our fosmids was checked by constructing phylogenetic trees of all the genes present in them. Despite the presence of several bacterial-like ORFs in some fosmids, the occurrence of critical housekeeping genes (for example ribosomal proteins, transcription factors, DNA polymerase subunits), typical archaeal functions (for example geranylgeranylglyceryl phosphate synthase), or consistent clustering with archaeal genes indicated that these fosmids were from Euryarchaeota of Groups II and III (we could detect only one ORF possibly related to Crenarchaeota). Many of the archaeal fosmids contained housekeeping genes such as ribosomal proteins (s28e, S10P, L7Ae, L24e), DNA topoisomerase IV, and transcriptional regulators and restriction enzymes that seemed to be closely related to *Aciduliprofundum boonei* (Reysenbach and Flores, 2008), a thermophilic archaeon isolated from a deep sea hydrothermal vent. This microbe seems to have the closest available genome to those of the Euryarchaeota Group II found in our sample. However, the microbes represented by our fosmids must have a very different lifestyle. One of the fosmids contained a rhodopsin gene along with a typically archaeal geranylgeranylglyceryl phosphate synthase gene (again most similar to *A. boonei*) (Figure 3). Rhodopsin genes enable widespread light-driven phototrophy in the open oceans (Beja *et al.*, 2000a). The rhodopsin gene in this case was not adjacent to a 16S rRNA gene, as was the case for the Group II Euryarchaeota fosmids as detected earlier in the HOT data set (Frigaard *et al.,* 2006). This may imply that the rhodopsin gene can be inserted in different genomic regions and/or that may be present in more than one copy in the genome. We found many reads annotated as rhodopsins in the DS sequences (39). Interestingly, a phylogenetic tree also indicated that some of the rhodopsins present cluster clearly with the type found in Euryarchaeota Group II (Figure 3). This finding was supported by high bootstrap values and also by the higher GC content of this clade and indicates that a significant fraction of the rhodopsin-based photoheterotrophy at the DCM could derive from members of the Euryarchaeota.

Some fosmids could be clearly assigned to certain bacteria, for example bacterium Ellin514 (3 fosmids), marine γ-proteobacterium HTCC2207 (1 fosmid), *Opitutus terrae* (Verrumicrobium, 1 fosmid), *Blastopirellula marina* DSM 3645 (1 fosmid), marine γ-proteobacterium HTCC2143 (1 fosmid), and others (total 33 fosmids). Only one fosmid was clearly ascribed to *Candidatus* Pelagibacter that from the 16S rRNA analysis of the DS should be by far the most abundant bacterium. This again illustrates the poor clonability of genomic fragments from this microbe (Temperton *et al.*, 2009).

Finally, a number of fosmids could be assigned to eukaryotic cells (24), with very low similarity, but some of them with genes consistently related to the diatom *Thalassiosira pseudonana* CCMP1335, and the green algae *Ostreococcus lucimarinus* CCE9901 confirming again the presence of these photosynthetic eukaryotes in the 5–0.2 μm fraction (see Supplementary File F2 for detailed annotation of the assembled fosmid sequences).



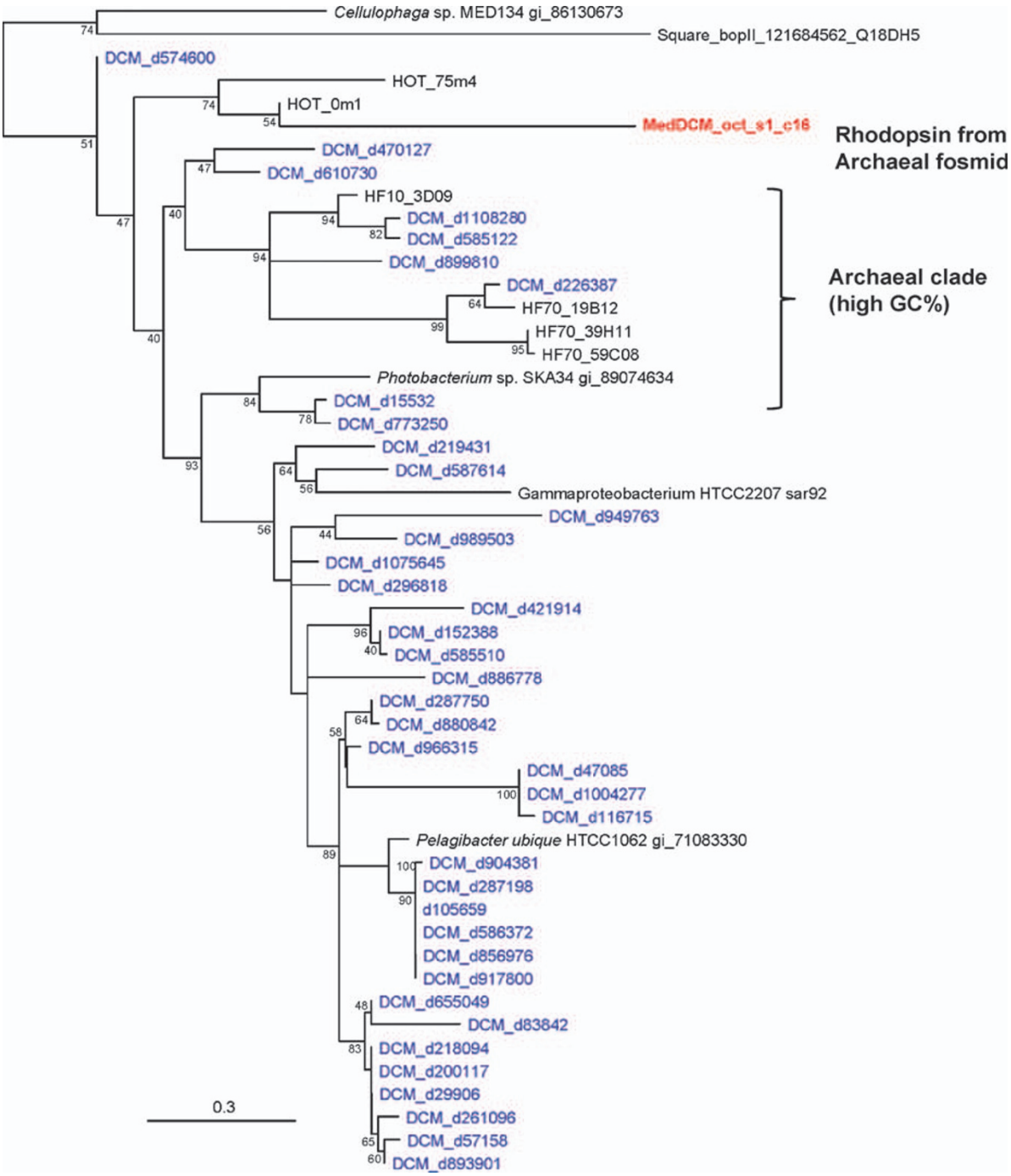


Figure 3 (Top) Archaeal fosmid containing geranylgeranylglyceryl phosphate synthase gene and proteorhodopsin genes (MedDCM-OCT-S08-C16). All unlabelled genes are hypothetical. (Below) Phylogenetic tree from all the rhodopsins identified in the DCM DS data set.

To assess the real significance of the microbes represented by the fosmids in the sample, we checked for recruitment at high identity levels in the DS data set (Supplementary Table T2). Fosmids that recruit well to the DS data are most likely to represent abundant microbes in the sample. The recruitment was actually quite uneven and there seemed to be no relationship with the GC content of the fosmid. Eukaryotic fosmids recruited very poorly, which is expected given the much larger size of eukaryotic genomes. Besides, the large number of eukaryotic fosmids found indicates that they are probably enriched in the cloning steps. The high clonability of eukaryotes and phages supports the idea that fosmid cloning favors DNA belonging to very different organisms from the *Escherichia coli* host, so that their expression and potential detrimental effect is minimized. On the other hand, several archaeal fosmids recruited quite well, confirming the presence of significant amounts of these microbes in the environment under study. Among the bacterial fosmids, the few retrieved belonging to *Candidatus* Pelagibacter and *P. marinus* were, as could be expected, recruiting very efficiently, but other fosmids with more uncertain and less expected affiliation also showed high recruitment. For example, the fosmid recruiting the maximum number of reads was of very mixed affiliation that could not be assigned to any specific phylogenetic group beyond Bacteria. The cyanophage fosmids recruited quite well (myoviruses more than podoviruses) (Supplementary Table T2). This confirms that these phages are actually a significant part of the biomass retrieved here.

Individual genome recruitment

A novel way to assess community structure is by recruitment analysis using the available genomes of microbes. The consistent recruitment at high identity levels can be considered the best proof of the presence of the microbe in the environment and if the identity is very high even infer the strain or ecotype to which the specific representative found in the metagenome belongs. We performed recruitment of the DCM DS data set sequences against all marine microbial genomes (complete or draft) available from NCBI. By far, the *P. marinus* subsp. *pastoris* strain CCMP1986 (MED4) was the genome that recruited the best. This particular microbe (originally referred to as *P. marinus* MED4) was first isolated in the Mediterranean Sea in 1989 at a depth of 5 m, and is a high-light-adapted strain (Rocap *et al*., 2003). Several other *Prochlorococcus* genomes also showed some amount of recruitment, but none were even close to the pastoris strain (Figure 4). The next highest recruiting genome was found to be of *Candidatus* Pelagibacter sp. HTCC7211. This strain was originally isolated from Bermuda in the Sargasso Sea. The other two *Candidatus* Pelagibacter genomes (HTCC1062 and HTCC1002), originally isolated near Oregon, in the colder waters of the Pacific (ca. 15 °C) recruited comparatively fewer reads. *Synechococcus* sp. WH 8102 was found to be

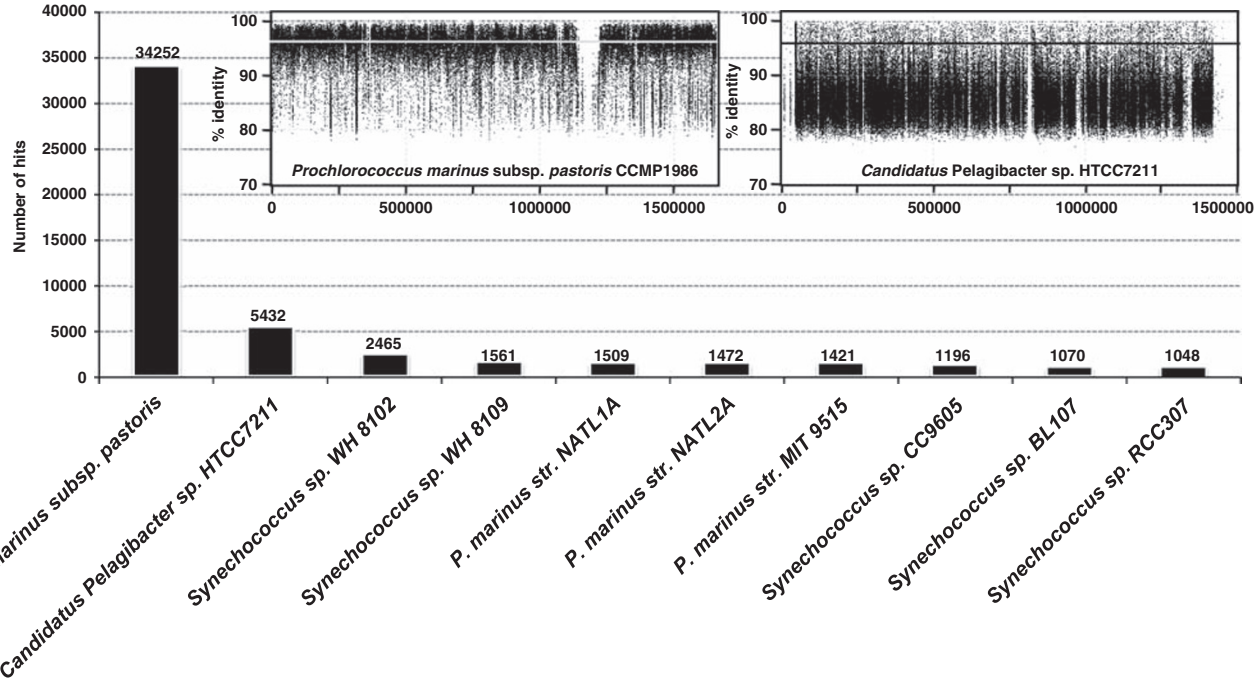


Figure 4 Recruitment of DS reads by complete genomes. Shown in the insets are the recruitment plots of the top two highest recruiting genomes *P. marinus* subsp. *pastoris* MED4 and *Candidatus* Pelagibacter sp. HTCC7211. The lines within the recruitment plots indicate the level of 96% nucleotide identity.

the third most abundant microbe by recruitment analysis. This particular strain was isolated in 1981 from the Atlantic Ocean. *Synechococcus* is usually less abundant, but much more broadly distributed than *Prochlorococcus*, which is found mostly within a latitude belt of ~48° N to 40° S (Scanlan *et al.*, 2009). Apart from these three, several other *Synechoccoccus* and *Prochlorococcus* genomes recruited a somewhat equal number of reads (Figure 4). Though the highest recruiting strain of *P. marinus* is high-light adapted, two low-light-adapted strains (NATL1A and NATL2A) were also found to recruit, but to a much lesser extent.

In addition to the microbial genomes, we also performed recruitment of the DCM DS sequences against all available marine phage genomes. The maximum number of reads were clearly recruited by cyanophages infecting *Prochlorococcus* and *Synechococcus* (see Supplementary Table T4), whereas other phage genomes, for example from *Roseobacter* phages or *Vibrio* phages, did not recruit any reads.

Physiology and ecology

We have compared our data sets with the GOS stations that cover a wide range of geographic and environmental sites, although all samples come from near the surface (Rusch *et al.*, 2007).

The GOS sample GS122a, collected in the Indian Ocean between South Africa and Madagascar, had the highest similarity to our sample (Figure 5). The next two closest samples were from the Sargasso Sea. The surface temperature at this site (20 °C) was similar to the average surface temperature in the Mediterranean. Although we do not have specific data, we might assume that the South African location also has a seasonal DCM, whereas most of the GOS samples come from tropical waters or the nutrient-rich cold waters of the Eastern North American seaboard. If that is the case, this similarity might be explained by the seasonal nature of the DCM at these locations. Furthermore, as the sample was taken during the austral wintertime (September), stratification would likely have been lost and the DCM dwellers will be found close to the surface.

The differential gene content that we found with GS122a could largely be ascribed to housekeeping archaeal genes (archaeal thermosome, archaeal flagellum) (see Supplementary File F1), reflecting the increased archaeal presence at the Mediterranean DCM. In addition, comparison of the DCM with the GOS surface samples (including Sargasso

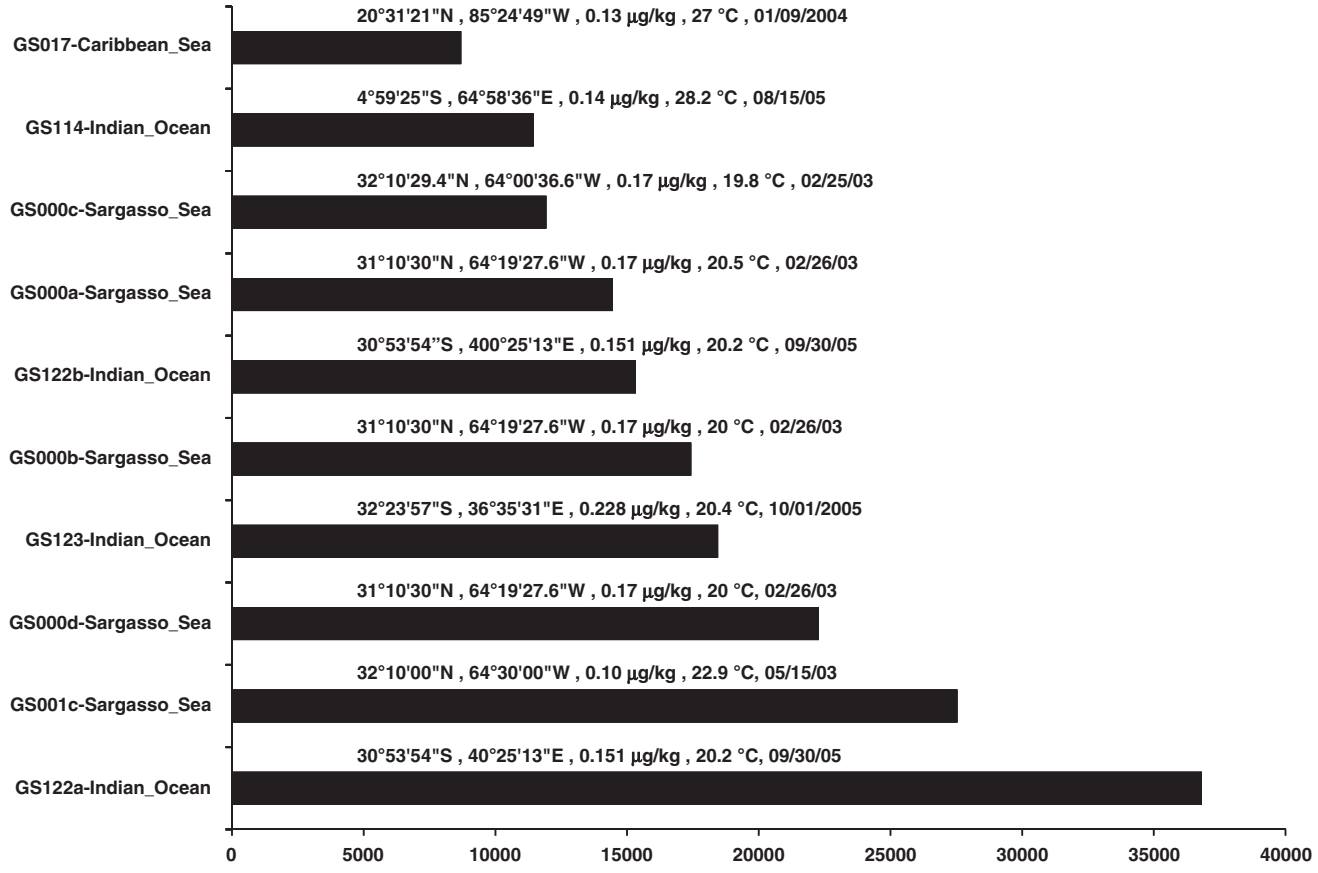


Figure 5 Comparison of DCM DS reads to GOS open ocean data set (minimum identity 95%, length 25 bp). The top 10 samples are shown. Shown above the bar for each sample are its latitude, longitude, chlorophyll concentration, temperature, and date of collection.

Sea samples) also indicated the increased archaeal presence at the DCM compared with the surface.

Interestingly, among the most overrepresented genes in the DCM compared with the surface were genes involved in phosphate transport and regulation. These included the periplasmic phosphate-binding protein pstC and other members of the phosphate-uptake regulon (PhoB, pstA, pstB). The Mediterranean is notorious for being phosphate limited and with higher ratio of inorganic nitrogen to phosphate (Krom *et al.*, 1991).

To analyze whether this indeed reflected the scarcity of phosphate, we compared our data set to several Sargasso Sea metagenomic data sets (phosphate poor), as well as several nutrient-rich GOS samples (Gulf of Maine, Bay of Fundy). The Sargasso GOS samples did not show any differences in the presence of phosphate-uptake-related genes, but the nutrient-rich samples gave a clear difference, similar to the comparison between GS122a and the DCM (Supplementary File F1). We did not find any differences in phosphonate-usage genes, as reported by earlier studies in the Eastern Mediterranean (Feingersch *et al.*, 2010).

A typical feature of the genomes of microbes from the aphotic zone is the absence of the *uvr* genes involved in DNA repair by photoreactivation. We did not find any difference between the number of *uvr* genes in the surface and the DCM. However, among the most overrepresented subsystems in the GOS surface samples compared with the DCM was the universal stress protein family. This is an ancient, widespread, and conserved family of proteins that are known to be expressed under different types of physiological stresses, such as heat stress, nutrient starvation, and DNA damage (Kvint *et al.*, 2003). The markedly inferior representations of these genes in the DCM community indicate that in spite of the annual mixing events, their dwellers live in a more constant environment, less affected by environmental fluctuations than would be at the surface.

Discussion

We have applied a similar 454 pyrosequencing effort to a sample from the Mediterranean seasonal DCM prior and after fosmid cloning, in an attempt to compare both methods to sequence metagenomic DNA, and also to get a glimpse at the community structure and ecosystem functioning of this important oceanographic feature. GC plots of the data sets revealed clear differences in the sequence obtained from both approaches (Figure 1). Two-peak GC plots have been described also for BAC libraries (Feingersch and Béjà, 2009), which was attributed to bias against SAR 11 (that includes *Candidatus* Pelagibacter). Besides, the bias against low-GC DNA has also been attributed to easier fragmentation during cloning (Temperton *et al.*, 2009). Several lines of evidence point toward the GC% distribution of the DS data set (Figure 1) being quite realistic. For example, earlier analysis of oligotrophic off-shore marine waters from GOS data (for example Sargasso Sea Hydrostation S) sequenced by the classic di-deoxy methodology gave very similar GC distribution plots. In addition, the GC content of the predominant microbes is concordant with these values. There is earlier evidence that indicates that as size ranges increase, a high-GC peak increases in relative abundance (Martin-Cuadrado *et al.*, 2008). In any case, it is possible that AT-rich sequences are overestimated as low-GC DNA seems to be better covered by 454 sequencing (Harismendy *et al.*, 2009). There is an additional bias in 454 sequencing that has been shown to result in artificial replicates, leading to overestimation of gene and taxon abundance (Gomez-Alvarez *et al.*, 2009). However, there seems to be general agreement about fosmid libraries discriminating against SAR 11 DNA. In our case, only one fosmid could be clearly classified as belonging to this group. However, some libraries, such as the one described before (Feingersch *et al.*, 2010), had many more clones belonging to this group. This might be due to slight modifications of the methodology (for example the use of BACs rather than fosmids), or because the sample was very enriched in such microbes. Still, even in the work of Feingersch *et al.*, SAR 11 were only 45% of the α-proteobacteria that were only about 42% of the total BAC ends (so only about 19% of the total end reads).

A very important outcome of metagenomic data sets is the possibility of assembling contigs belonging to the same microbe. In the case of the fosmid library, the potential for assembly has to be larger, at least in a high complexity sample such as the DCM. The 1152 fosmids that were pooled for sequencing should contain between 30 and 40 Mb of natural DNA, so we should have achieved nearly 10 × coverage and large contigs were expected. However, as shown in Supplementary Table T1, only about 8 Mb could be assembled into contigs of >3 kb. There was considerable variation in the amount of sequence obtained from 454 sequencing of each sample (see Supplementary Table T5; each sample had DNA from 96 pooled fosmids, and there were a total of 12 samples). As the fosmids were pooled together before the DNA was extracted, it is likely that, even though a copy control vector was used, there might be small differences in the DNA amounts obtained from each fosmid. This combined with the varying amount of sequence obtained from each sample affects real coverage achieved for each fosmid, thus affecting assembly. This is clearly also reflected in widely varying values for average fosmid coverage among the longest-assembled contigs (see Supplementary Figures S3 and S4). A smaller number of fosmids per 454 plates might have provided better assembly. Still, the number of contigs >10 kb was close to 200 representing >3 Mb, and all had high coverage (>10 ×). In

Table 1 Performance of sequencing of metagenomic fosmids versus DS

Analysis	*Fosmid sequencing*	*Direct sequencing*
Total environmental sequence generated	325 Mb	312 Mb
rRNA reads	100	1058
Number of assembled contigs (>10 kb)	197 (longest contig: 44 kb)	5 (longest contig: 4.4 kb)
Bias	Cloning bias against SAR11 and Cyanobacteria	Bias because of artificial replicates during emPCR and bias against high GC%
Taxonomic assignment	Highly reliable	Unreliable for underrepresented genomes (e.g. Archaea)
Functional insights	Easy to infer (e.g. archaeal rhodopsin)	Only possible for general ecological comparisons (e.g. phosphate transport)

Abbreviations: DS, direct sequencing; rRNA, ribosomal RNA.

contrast, <1 Mb could be assembled from DS data set. In addition, the assembly of long DS contigs was restricted only to the highly predominant genomes, whereas the fosmids provided contigs belonging to a wide diversity of microbes. The cloning bias detected here favoring fosmid inserts with skewed GC content has also been identified in shotgun libraries of different sample types (Sorek *et al*., 2007) and adds to the logical bias against genes that may be toxic to the host. These factors imply an important constraint to metagenomic studies and, therefore, the use of different hosts with a different GC content as well as the usage of cloning vectors where the DNA inserts can be transcriptionally silenced would be highly desirable. As a methodological strategy, we recommend the use of pyrosequencing of total metagenomic DNA as well as of fosmids to get a more complete picture of the structure and composition of natural prokaryotic communities, as both approaches seem to be complementary (Table 1).

For example, this combination of approaches provides a robust description of the community structure. Our data show a remarkable number of similarities with other DCM descriptions from around the world. The dominance of picocyanobacteria is well known and has been described by different techniques (for example flow cytometry) even before molecular approaches were available. In the study carried out using fosmids in the HOT DCM (70 m), picocyanobacteria accounted for about 10% of the fosmid end sequences (DeLong *et al*., 2006). However, the massive local dominance of a specific strain of *P. marinus* subsp. *pastoris* (MED4), a high-light-adapted ecotype (Rocap *et al*., 2003), in the microbial community at such extremely high-sequence identity (compared with other high- or low-light ecotypes) came as a surprise. It seems clear that close relatives of this specific strain make up a large fraction of the photosynthetic population that we have studied. This is not only borne out mostly by the genome recruitment observations, but also the largest contig directly assembled from the DS belongs was totally syntenic and had a high similarity to the genome of MED4. The dominance of the *P. marinus* MED4 ecotype (identified by the *pcb* gene) in the Mediterranean had been suggested before (Garczarek *et al*., 2007). Contrastingly, although we found many cyanophages that had similar genomes to those described for picocyanobacteria, the similarity that we found was never so high. This may reflect a higher variability among the phages than among their host or simply a poorer availability of cyanophage genomes. The fact that MED4 was isolated from the Mediterranean (although 800 km away from our sampling location) might partially explain this high conservation. However, this genome recruited quite well also in the Indian ocean GOS samples (data not shown), so it might be a reflection of habitat selection rather than geography. We did obtain some amount of recruitment with the the low-light-adapted ecotypes of *P. marinus* (for example NATL1A and NATL2A) genomes as well, but they are all from different geographic areas, so it is possible that they are different in the Mediterranean and have not been sequenced.

Cyanophages must constitute a significant fraction of the genetic material in the DCM. For example, cyanophage fosmids constitute nearly 10% of the long-assembled fosmid sequences, several of these fosmids recruit well, and some of the few long contigs assembled from the DS reads belonged to cyanophages. We expected most viruses to escape and not be collected in the filter size, but a similar situation was described for the HOT photic zone, a sample that was processed similarly. Phages were more predominant at 70 m depth (the beginning of the DCM), and cyanophages accounted for 12% of all sequences and there was a relative underrepresentation of cyanobacterial cellular DNA (DeLong *et al*, 2006). An explanation is that at the time of collection, the phages were replicating in the cyanobacterial cells and have been captured within the cellular fraction as has also been earlier suggested (DeLong *et al*., 2006). If this is a regular feature of this habitat, it strengthens the idea that viruses are vital regulators of bacterial population dynamics (Brockhurst *et al*., 2006; Rodriguez-Valera *et al*, 2009).

Another interesting aspect regarding the community structure is the relatively high abundance of Euryarchaeota in the sample and specifically of a group that has genomic level relationships to the

thermophile *A. boonei*. Both the presence of this group in the tropical DCM and the presence of rhodopsins in some representatives were already known and are confirmed in this study. The increased archaeal presence at the DCM was also supported by the comparisons with the GOS surface samples. It is important to stress that these microbes seem to be very scarce in the upper regions of the photic zone. Efforts should be directed to isolate pure cultures or at least to collect sequence data from these microbes at DCM and other deep photic zone habitats.

Acknowledgements

This work was supported by projects GEN2007-30014E, BIO2008-02444, and CONSOLIDER Ingenio 2010 CSD2009-00006 and PROFIT 170/07. A.-B.M.-C. was supported by a Juan de la Cierva scholarship, all from the Spanish Ministerio de Ciencia e Innovación, and ACOMP2009/359 from the Generalitat Valenciana. PLG, PD, and DM were supported by the French National Agency for Research (EVOLDEEP project, contract number ANR-08-GENM-024-002). AM and RC were funded by the Explora Project BIO2008-03419-E from the Spanish Ministerio de Ciencia e Innovación.

References

Altschul SF, Madden TL, Schäffer 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.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al.* (2000a). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902–1906.

Beja O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP *et al.* (2000b). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516–529.

Brockhurst MA, Fenton A, Roulston B, Rainey PB. (2006). The impact of phages on interspecific competition in experimental populations of bacteria. *BMC Ecol* **6**: 19.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–D145.

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.

Edgar RC. (2004). MUSCLE: a multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **5**: 1792–1797.

Estrada M, Marrase C, Latasa M, Berdalet E, Delgado M, Riera T. (1993). Variability of the deep chlorophyll máximum in the Northwestern Mediterranean Sea. *Mar Ecol Prog Ser* **92**: 289–300.

Feingersch R, Béjà O. (2009). Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries? *ISME J* **10**: 1117–1119.

Feingersch R, Suzuki MT, Shmoish M, Sharon I, Sabehi G, Partensky F *et al.* 2010. Microbial community genomics in eastern Mediterranean Sea surface waters. *ISME J* **4**: 78–87.

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

Garczarek L, Dufresne A, Rousval S, West NJ, Mazard S, Marie D *et al.* (2007). High vertical and low horizontal diversity of *Prochlorococcus* ecotypes in the Mediterranean Sea in summer. *FEMS Microbial Ecol* **60**: 189–206.

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

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY *et al.* (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: R32.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **3**: 377–386.

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

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* **16**: 5345–5355.

Krom MD, Kress N, Brenner S. (1991). Phosphorus limitation of primary productivity in the eastern Mediterranean Sea. *Limnol Oceangr* **3**: 424–432.

Kvint K, Nachin L, Diez A, Nyström T. (2003). The bacterial universal stress protein: function and regulation. *Curr Opin Microbiol* **6**: 140–145.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. (2004). Transfer of photosysnthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* **101**: 11013–11018.

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.

Martín-Cuadrado AB, López-García P, Alba JC, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic hábitat. *PLoS One* **9**: e914.

Martin-Cuadrado AB, Rodriguez-Valera F, Moreira D, Alba JC, Ivars-Martínez 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* **8**: 865–886.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Noguchi H, Taniguchi T, Itoh T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* **6**: 387–396.

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* **21**: 7188–7196.

Reysenbach AL, Flores GE. (2008). Electron microscopy encounters with unusual thermophiles helps direct genome analysis of Aciduliprofundum boonei. *Geobiology* **3**: 331–336.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F *et al.* (2009). Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* **11**: 828–836.

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* **3**: e77.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **2**: 249–299.

Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449–1452.

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **8**: e234.

Temperton B, Field D, Oliver A, Tiwari B, Mühling M, Joint I *et al.* (2009). Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J*, 2009 **7**: 792–796.

Vaulot D, Eikrem W, Viprey M, Moreau H. (2008). The diversity of small eukaryotic phytoplankton ($<$ or $=3$ microm) in marine ecosystems. *FEMS Microbiol Rev* **32**: 795–820.

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.

Zaballos M, López-López A, Ovreas L, Bartual SG, D'Auria G, Alba JC *et al.* (2006). Comparison of prokaryotic diversity at offshore oceanic locations reveals a different microbiota in the Mediterranean Sea. *FEMS Microbiol Ecol* **3**: 389–405.

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