The ISME Journal (2010) 4, 78–87

www.nature.com/ismej

ORIGINAL ARTICLE

Microbial community genomics in eastern Mediterranean Sea surface waters

Roi Feingersch[1], Marcelino T Suzuki[2,6], Michael Shmoish[3], Itai Sharon[1,4], Gazalah Sabehi[1], Frédéric Partensky[5] and Oded Béjà[1]

[1]*Faculty of Biology, Technion—Israel Institute of Technology, Haifa, Israel;* [2]*Chesapeake Biological Laboratory, University of Maryland Center of Environmental Science, MD, USA;* [3]*Bioinformatics Knowledge Unit, Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering, Technion—Israel Institute of Technology, Haifa, Israel;* [4]*Computer Science Department, Technion—Israel Institute of Technology, Haifa, Israel and* [5]*CNRS and Université Pierre et Marie Curie (Paris 6), UMR 7144, Station Biologique, Roscoff, France*

Offshore waters of the eastern Mediterranean Sea are one of the most oligotrophic regions on Earth in which the primary productivity is phosphorus limited. To study the unexplored function and physiology of microbes inhabiting this system, we have analyzed a genomic library from the eastern Mediterranean Sea surface waters by sequencing both termini of nearly 5000 clones. Genome recruitment strategies showed that the majority of high-scoring pairs corresponded to genomes from the *Alphaproteobacteria* (SAR11-like and *Rhodobacterales*), *Cyanobacteria* (*Synechococcus* and high-light adapted *Prochlorococcus*) and diverse uncultured *Gammaproteobacteria*. The community structure observed, as evaluated by both protein similarity scores or metabolic potential, was similar to that found in the euphotic zone of the ALOHA station off Hawaii but very different from that of deep aphotic zones in both the Mediterranean Sea and the Pacific Ocean. In addition, a strong enrichment toward phosphate and phosphonate uptake and utilization metabolism was also observed.
The ISME Journal (2010) **4,** 78–87; doi:10.1038/ismej.2009.92; published online 20 August 2009
Subject Category: integrated genomics and post-genomic approaches in microbial ecology
Keywords: metagenomics; sar11; phosphonate; Mediterranean Sea

Introduction

The marine environment covers approximately 71% of the Earth's surface and contains an enormous pool of diverse microbes occupying a variety of niches. Although these marine microbes account for very large fluxes of energy and matter, which influence the whole biosphere, their genetic and metabolic diversity are just starting to be fully explored (DeLong *et al.*, 2006; Pommier *et al.*, 2007). The recent introduction of genomic tools has revolutionized our view of microbial diversity and activity, and has shown that our knowledge of natural environments has been highly biased, as a result of the earlier focus on cultured microbes, which represent less than 1% of existing marine organisms (Kogure *et al.*, 1979; Rusch *et al.*, 2007). Cultivation-independent methods such as gene surveys, whole-genome shotgun sequencing and large-insert DNA libraries (such as bacterial artificial chromosomes (BACs) and fosmids) enable us to bypass cultivation biases and study microorganisms directly from the environment (for a review, see DeLong, 2005). Molecular approaches based on the amplification of ribosomal RNA (rRNA) genes, although allowing us to tackle the phylogenetic diversity of indigenous microbes, do not always provide insights regarding the metabolic and functional potential of microorganisms, and do not account for substantial differences in genome size and content that can occur between individuals from the same species (Béjà *et al.*, 2000; Pedros-Alio, 2006). Organisms that share more than 97% identity in their 16S rRNA sequence, which are commonly classified as members of the same species, may show wide differences in the physiology, biochemistry and genome content, which in turn can be shared with other species that inhabit the same ecosystem (Sikorski and Nevo, 2005; Rusch *et al.*, 2007). DeLong *et al.* (2006) were the first to combine the construction of large-insert bacterial fosmid libraries with shotgun sequencing of clone termini. The benefit of such large DNA inserts is that they may contain operons and sometimes even entire pathways, providing a way to examine metabolic potentialities and gene organization in field microorganisms (Béjà, 2004).

Correspondence: O Béjà, Faculty of Biology, Technion—Israel Institute of Technology, Haifa 32000, Israel.
E-mail: beja@techunix.technion.ac.il
[6]Current address: CNRS and Université Pierre et Marie Curie (Paris 6), Observatoire Océanologique de Banyuls, BP44, Banyuls-sur-Mer, France.
Received 18 May 2009; revised 13 July 2009; accepted 14 July 2009; published online 20 August 2009

In regard to inorganic nutrients, offshore waters of the eastern Mediterranean Sea are one of the poorest oceanic regions on Earth (Krom *et al.*, 1991). As a result of these ultraoligotrophic conditions, the water is very clear and there are low rates of primary production and a dominance of small-sized phytoplankton (Thingstad *et al.*, 2005). The few currently available studies on the prokaryotic community composition of the water column of the eastern Mediterranean Sea (Moeseneder *et al.*, 2001a, 2001b, 2005) all studied deep waters (deeper than 200 m). Here, we report the analysis of BAC ends from a metagenomic library constructed from the eastern Mediterranean Sea surface water.

Materials and methods

Sample collection and BAC library construction
The BAC library was constructed from picoplankton samples from 12 m deep water collected on a transect from Haifa to Cyprus (33°25′N, 33°56′E to 32°54′N, 34°44′E) during August 2003 (Sabehi *et al.*, 2005). BAC library construction was carried out as described earlier (Béjà *et al.*, 2000; Sabehi and Béjà, 2007). Approximately 800 l were pre-filtered through Whatman GF/A filters (Middlesex, UK), (1.6 μm equivalent pore size) to remove most of the larger eukaryotic phytoplankton cells. Bacterioplankton were concentrated by tangential flow filtration with a Pellicon-2 unit (Millipore, Billerica, MA, USA) equipped with a type C 30 polysulfone membrane cartridge (30 000 MW cutoff and 0.5 m^2 cassette) (Biomax, Martinsried, Germany). Bacterioplankton cells were collected by centrifugation of the retentate (4 °C, 38 900 *g*, 1 h). The bacterioplankton pellet was embedded in agarose plugs, and DNA was extracted and cloned into pBACIndigo536, a derivative of the pBeloBAC11 (Kim *et al.*, 1996) vector. Out of the 10 176 BAC-end sequences, 7035 were used for further analyses after quality control and removal of duplicates.

Community composition and taxonomic classification
Partial 16S and 23S rRNA gene sequences were identified using a multistep strategy. Whenever possible, 16S rRNA sequences were identified using the online ribosomal database project (release 10) naive bayesian classifier using a confidence threshold of 95% (Cole *et al.*, 2009). Sequences not classifiable at the genus level by the RDP classifier or where a more detailed classification was possible were identified by blast searches (megablast, no filters) against the GenBank 'nt' database using blastcl3 and classified on the basis of matches with greater than 99% identities over 99% of the queried sequences. Queried sequences were identified into clades on the basis of placement of matching sequences in phylogenetic trees published earlier (Rappé *et al.*, 1997; Suzuki *et al.*, 2001, 2004; Rocap *et al.*, 2002; Kan *et al.*, 2008) and the greengenes (DeSantis *et al.*, 2006) classification.

23S rRNA gene sequences were classified using blast searches (megablast, no filters) against the GenBank 'nt', as well as the 'Microbial/408172_13694' database containing assemblies from the Global Ocean Sampling study (Rusch *et al.*, 2007) and with one exception classified on the basis of matching sequences with greater than 98% identities to a fragment spanning over 99% of the queried sequences. Whenever present, the top 23S hits were identified on the basis of syntenic 16S rRNA as described above or by importing these 16S rRNA sequences into an arb database described earlier (Kan *et al.*, 2008), aligning and adding these sequences to trees published earlier by arb_parsimony as described earlier (Kan *et al.*, 2008).

Protein-based community composition and taxonomic classification were made by BLASTx querying (-p blastx -e 1e-5 -F F) against the NCBI non-redundant (nr) database with an expectation value of $\leqslant 1 \times 10^{-50}$. The top high-scoring sequence pair of a query was determined according to the NCBI taxonomic identifier. For cross-classification, the BAC termini were classified with an expectation value $\leqslant 1 \times 10^{-10}$.

Fragment recruitment
The Mediterranean BAC library was aligned using the NCBI BLASTn. The recruitment was made with fragments of more then 200 bp, expectation value of $\leqslant 1 \times 10^{-5}$ (-p blastn -e 1e-5 -F F -r 2) and the BLASTn high-scoring sequence pair cover more than 80% of the hit. Data were plotted with the Gnuplot program (http://www.gnuplot.info).

Metabolic potential
Metabolic potential was determined with the integrated microbial genome (Markowitz *et al.*, 2008) of the Joint Genome Institute server (http://img.jgi.doe.gov/m). Frequency estimations of the different protein families were computed on the basis of the statistical framework from Sharon *et al.* (2009). Briefly, the frequency F_P of a protein family P in a sample D is given by

$$F_P = \frac{\hat{C}_P}{\sum_{Q \in D} \hat{C}_Q}$$

where $\hat{C}_P$ is an estimator for the number of occurrences of P in the sample, $\hat{C}_P$. Given that P is observed on R_P reads, $\hat{C}_P$ is given by

$$\hat{C}_P = \frac{R_P}{2\alpha(L(P) + g - 2T)}$$

where *L(P)* is the length of *P*, *g* is the average read length, *T* is the minimum segment of *P* required for the gene to be discovered and α is the sample coverage. Note that α is eliminated when F_P is computed.

The Cluster of Orthologous Group (COG) assignment of the nine environments was tabulated and can be found in Supplementary Table S1. We used the Expander 4.0.2 for clustering (Shamir *et al.*, 2005). From all the COGs, we chose the 1000 most variable ones, performed standardization (mean 0, variance 1 transformation) and a hierarchical clustering with complete linkage (Figure 4b).

Library-to-library similarities
Measurement of library-to-library similarity was determined after BLASTn querying of all environments against each other and themselves (-p tblastx -e 1e-2 -F F). The data set was computed according to Rusch *et al.* (2007) with a cutoff value of 80% and only for fragments of more than 200 bp (Figure 4a).

Nucleotide sequence accession numbers
End sequences were deposited in GenBank under accession numbers GQ233105–GQ240138.

Results and discussion

General features of the BAC library
To gain information on the composition and structure of the microbial community in the eastern Mediterranean surface water, 10 176 randomly sequenced BAC ends from an eastern Mediterranean BAC library (Sabehi *et al.*, 2005) were analyzed. This BAC library was constructed from picoplankton samples from 12 m deep water collected on a transect from Haifa to Cyprus (33°25′N, 33°56′E to 32°54′N, 34°44′E) during August 2003 and prefiltered through a GF/A (1.6 μm equivalent pore size) glass fiber filters (Sabehi *et al.*, 2005).

Out of the 10 176 BAC-end sequences, 7 035 were used for further analyses after quality control and removal of duplicates. The BAC ends yielded approximately 5.4 Mb of DNA sequences from about 800 Mb archived in the library.

Community composition
Bacterial phylogenetic distributions based on 16S and 23S rRNA genes were generally consistent with earlier PCR-based cultivation-independent rRNA surveys of marine picoplankton (Rappé and Giovannoni, 2003). Fifty-three BAC ends containing rRNA genes (16S or 23S) were identified (Table 1 and Supplementary Table S1). They included representatives from *Synechococcus* clades II, III and IV (Fuller *et al.*, 2003); *Prochlorococcus marinus* high light (HL)-adapted clades I and II (West and Scanlan, 1999; West *et al.*, 2001); *Verrucomicrobiales*; *Gammaproteobacteria* OM60/NOR5, CHAB-I-7, uncultured RedeBac7D11 (Kan *et al.*, 2008) and SAR92 and SAR86-I clades (Suzuki *et al.*, 2001; Cho and Giovannoni, 2004); *Alphaproteobacteria* from the uncultured *Rhodospirillales* SPOTSAUG01_5m94 clade and SAR11 IB subclades (Kan *et al.*, 2008); and uncultured marine *Actinobacteria* belonging to the OM1 clade, a planktonic clade originally reported by Rappé *et al.*, 1997.

Table 1 Phylogenetic classification of BAC-end 16S and 23S rRNAs by the scheme described in the text. Number of reads associated to each phylum/class is the sum of reads belonging to its clades/subclades

Phylogenetic identification phylum/class, clade, subclade	*Number of reads*	*Subclade reference*
Actinobacteria	1	
OM1	1	Rappé *et al.* (1997)
Alphaproteobacteria	6	
Rhodospirillales		
SPOTSAUG01_5m94	2	Kan *et al.* (2008)
SAR11		
SAR11-IB	1	Suzuki *et al.* (2004)
Unidentifiable	2	
Cyanobacteria	36	
Prochlorococcus		
Low B/A clade I	1	Rocap *et al.* (2002)
Low B/A clade II	2	Rocap *et al.* (2002)
Synechococcus		
Clade II	24	Rocap *et al.* (2002)
Clade III	6	Rocap *et al.* (2002)
Clade IV	1	Rocap *et al.* (2002)
Unidentifiable (short)	1	
Unidentifiable	1	
Gammaproteobacteria	9	
CHAB-I-7	1	Kan *et al.* (2008)
RedeBac7D11	2	Kan *et al.* (2008)
SAR86		
Clade I	2	Suzuki *et al* (2001)
Unidentifiable	5	
Verrucomicrobia	1	
Opitutaceae		
Opitutus	1	RDP classifier

We used BLASTx searches with expectation values of $\leqslant 1 \times 10^{-50}$ to bin our BAC-end sequences within known taxa with sequences in gene databases. According to BAC-end taxon-binning, about 42% of the microbial BAC ends in our Mediterranean Sea library belonged to the *Alphaproteobacteria*, followed by *Cyanobacteria* that accounted for approximately 28% of sequences and *Gammaproteobacteria*, which represented less than 13%. *Bacteroidetes/Chlorobi* and Gram-positive bacteria (*Firmicutes* and *Actinobacteria*) sequences were also found in the library, yet with lower frequency (~2.5% each).

Not surprisingly, on the basis of this analysis, the abundant and cosmopolitan SAR11 group (Morris *et al.*, 2002) represented ~45% of the *Alphaproteobacteria*, whereas the *Rhodobacterales* were the next most common *Alphaproteobacteria* group (~26%). These estimates fit earlier abundance analyses of the SAR11 clade in the Mediterranean coastal waters during the summer (~20% of 4,6-diamidino-2-phenylindole counts; Alonso-Sáez *et al.* (2007). It is noteworthy that only one BAC end (out of 53

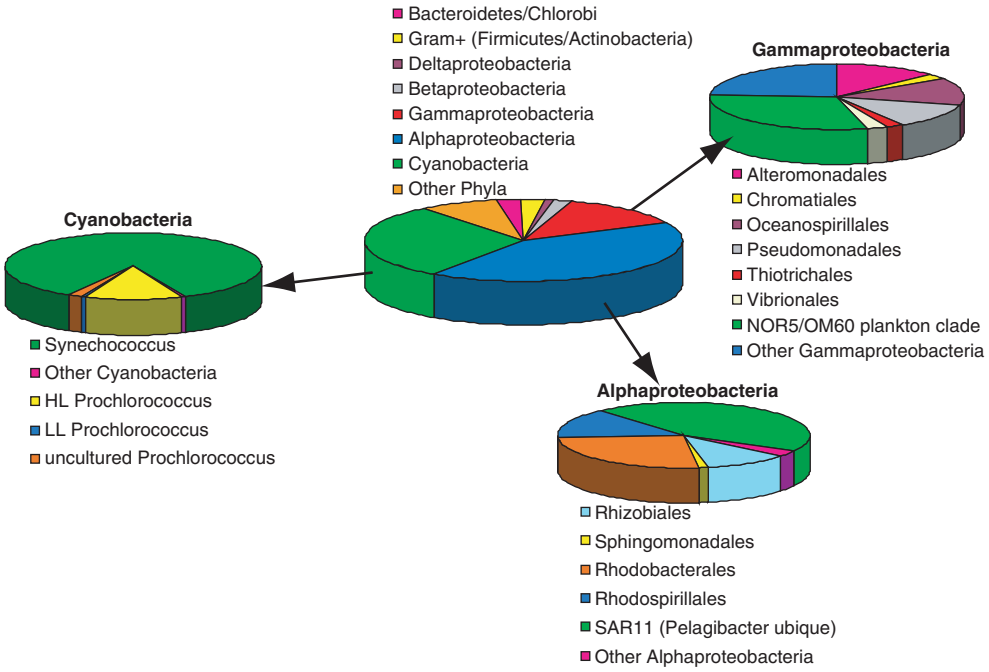

Figure 1 Prokaryotic taxa distribution in the Mediterranean bacterial artificial chromosome library inferred from end sequencing. The size of each taxon is relative to the number of top high-scoring sequence pair with expectation values of $\leqslant 1 \times 10^{-50}$. The category 'uncultured *Prochlorococcus*' refers to fosmid sequences from ALOHA surface waters (Coleman *et al.*, 2006).

rRNA-containing ends) was affiliated with SAR11. This discrepancy between low proportions of BAC ends containing SAR11 rRNA genes (1.9%) and the high proportion (~18%) of SAR11 seen in the high-scoring sequence pairs is consistent with earlier reports from surface stations of the ALOHA community genomics fosmid project (DeLong *et al.*, 2006; Pham *et al.*, 2008).

The BAC ends classified as *Cyanobacteria* were mainly composed of *Synechococcus* (~85%) and HL *Prochlorococcus* sequences (~12%). Although no cell counts are available for the samples used in this study, flow cytometric measurements have been reported from nearby stations of the Mediterranean Sea sampled in May 2001 and 2002 by Tanaka *et al.* (2007). These researchers showed that there were about $2–4 \times 10^3$ *Synechococcus* ml^{-1} in surface waters, whereas *Prochlorococcus* cells were present at very low concentrations above 50 m depth. Although this may be due in part to the low fluorescence of *Prochlorococcus* cells that makes them difficult to detect by flow cytometry in the upper water column, our metagenomic data are consistent with the fact that *Prochlorococcus* cells were indeed at significantly lower concentrations than *Synechococcus* in our samples, even taking into account differences in genome sizes between the two cell types (~1.7 vs 2.5 Mb for HL *Prochlorococcus* and for *Synechococcus*, respectively (Kettler *et al.*, 2007; Dufresne *et al.*, 2008). An earlier study on the diversity of *Prochlorococcus* in the Mediterranean Sea using *in situ* hybridization and probes specific for the HL and low light-adapted (hereafter LL) ecotypes showed a clear-cut vertical partitioning of these two groups, associated with the strong vertical stratification typical for Mediterranean Sea waters in summer (Garczarek *et al.*, 2007). Thus, it was somehow unexpected that a small fraction of our BAC-end sequences had a best match to LL-adapted *Prochlorococcus* strains (Figure 1). Nevertheless, Johnson *et al.* (2006) found that the cells of one LL clade (called 'eNATL2A' or 'LLI') could be retrieved up to the surface, but only when waters were vertically mixed, in particular at high latitude or in upwelling waters.

The *Gammaproteobacteria* BAC ends were mainly represented by the NOR5/OM60 group (30%). This clade includes the cultured *Congregibacter litoralis* strain KT71 (Fuchs *et al.*, 2007) and strain HTCC2080 (Cho *et al.*, 2007), the first cultured representatives of marine aerobic anoxygenic phototrophic *Gammaproteobacteria.* This clade was recently shown to have a cosmopolitan occurrence in the marine environment, with a clear preference for coastal waters (Yan *et al.*, 2009).

There might be some biases in the relative proportion of these phyla because of the fact that, in spite of an increasing effort to widen the taxonomic spectrum of fully sequenced genomes in gene databases, there are many more genomes available from *Proteobacteria* and *Firmicutes* than from other taxa, which might result in an overrepresentation of these lineages (Martín-Cuadrado *et al.*, 2007). Variations of genome sizes can also affect the outcome of our analysis. Proteins with known homologs may also have ambiguous taxonomic classification and have been erroneously placed in a phylum with relatively low scores as a result of the absence of closely related genomes in sequence databases.

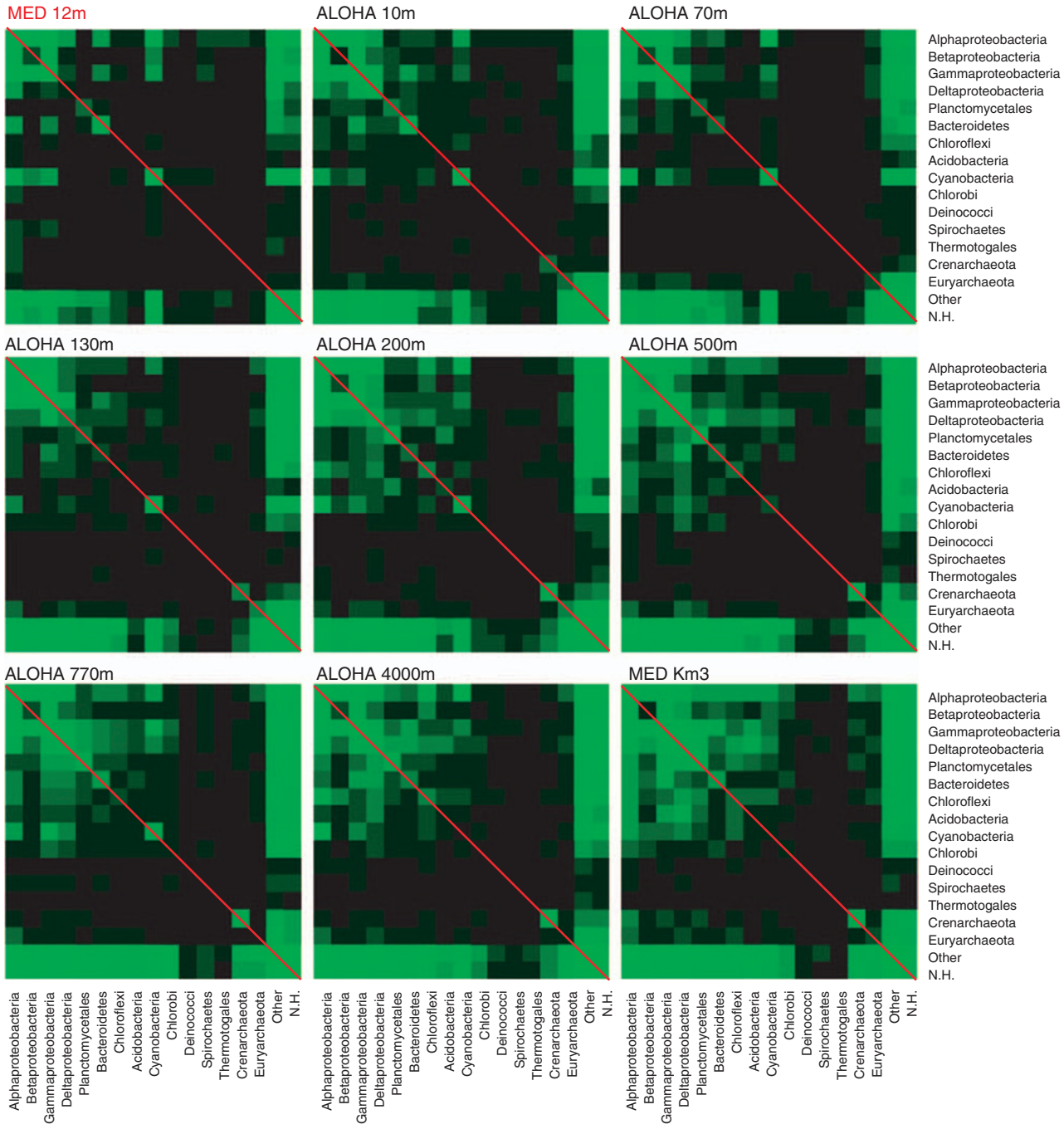

Figure 2 Cross-classification of each of the end sequences of bacterial artificial chromosomes or fosmid clones from nine different environments. Each end was classified as in Figure 1 with BLASTx expectation values lower than 1×10^{-10} into 15 major taxa. 'Other' stands for a taxon not belonging to any of those categories. 'N.H.' represents an end that did not qualify the classification terms. The intensity of the green shade is proportional to the percentage of ends with the same cross-classification. 'ALOHA' represents the different fosmid libraries from the ALOHA station (DeLong *et al.*, 2006); 'MED Km3' represents a Mediterranean fosmid library from 3000 m (Martín-Cuadrado *et al.*, 2007). A full colour version of this figure is available at *The ISME Journal* online.

This phenomenon of ambiguous taxonomic classification is illustrated in Figure 2, which shows how one end of a BAC end can be ascribed to a given taxon, whereas the other end is attributed to a different taxon. For example, in our surface Mediterranean library (MED 12 m in Figure 2), it is visible that although BAC inserts associated with *Alphaproteobacteria* are most common, there is a significant amount of BAC inserts that are attributed to *Alphaproteobacteria* at one BAC end and a different taxon group at the other end, such as *Betaproteobacteria*, *Gammaproteobacteria*, *Cyanobacteria* or *Bacteroidetes*. Figure 2 also shows that samples from deeper in the water column are characterized

by a broader distribution of ends not corresponding to the same taxon, that is, not found on the (red) oblique line. This might reflect a lateral gene transfer, as suggested earlier by Nesbø *et al.* (2005). However, as most fully sequenced genomes (~80%) still heavily represent only three bacterial phyla (that is, *Proteobacteria*, *Firmicutes* and *Actinobacteria*), because of their more ready culturability (Hugenholtz and Kyrpides, 2009; Kyrpides, 2009), we tend to favor an alternative interpretation to our results. We believe that the discordance between BAC-end identities is because of the lack of representation of deep oceanic representatives in bacterial culture collections, and consequently in fully sequenced genome databases. This way of representing genomic sequence data also allows a quick identification of taxa that are particularly well represented in certain environments (for example, the *Betaproteobacteria* found in ALOHA 130 m or the *Crenarchaeota* observed only in deep water stations below 200 m).

DNA fragment recruitment

An additional way of estimating microbial composition or diversity is by recruiting individual reads from environmental genomic libraries onto known microbial genomes (Rusch *et al.* 2007). We recruited our BAC-end sequences on a variety of *Alphaproteobacteria*, *Cyanobacteria* and *Gammaproteobacteria* genomes. As shown in Figure 3, for *Alphaproteobacteria* genomes, only SAR11 ('*Candidatus* Pelagibacter ubique') genomes yielded a fair recruitment of the BAC ends, although identities in general did not exceed 85% (DNA level). This probably points to the fact that our library is composed of yet to be cultured members of the SAR11 clade. Interestingly, recruitment was very low onto *Rhodobacterales* genomes.

Recruitments of a moderate number of BAC ends (Figure 3) at high similarity levels (>95%) to the genome of *Synechococcus* sp. WH8102 (clade III) and of a larger number of fragments at a lower similarity level (90–95%) to *Synechococcus* sp. CC9605 (clade II) suggest that the *Synechococcus* community was likely composed in a majority of members of clades II and III (*sensu* Fuller *et al.*, 2003). Generally speaking, a number of molecular studies using the hybridization of probes targeting the different *Synechococcus* clades have suggested that clade III is confined to oligotrophic areas, whereas clade II preferentially thrives in warm, coastal or continental shelf areas (Zwirglmaier *et al.*, 2007). A study of the diversity of *Synechococcus* off the coast of California furthermore showed that clade II populations are restricted to the upper mixed layer (Toledo and Palenik, 2003). More specifically, Mazard (2007) studied the diversity of *Synechococcus* in the Mediterranean Sea in summer along a transect from Gibraltar to a station off the coast of Libya and back to the south coast of France (PROSOPE cruise, summer 1999); she showed that clade III dominated in the oligotrophic waters of the Eastern basin, whereas clade II represented only a small percentage of the *Synechococcus* community at all stations. In contrast, in the Red Sea, clade II was found to predominate in the community all year round except for a peak of clade III in June (Fuller *et al.*, 2005). Finally, our BAC-end sequences showed only weak similarity to representative strains of clades I and IV, which are known to co-dominate in cold, mesotrophic waters. Altogether, our data are consistent with earlier molecular analyses, given the fact that our initial sample was a mixture of surface samples collected in warm surface waters along a coast-offshore transect.

The results are more surprising for *Prochlorococcus*. Indeed, BAC-end sequences are significantly closer from the HLII strain AS9601 than from the HLI strain MED4 (Figure 3). This is at odds with previous analyses of *Prochlorococcus* diversity in the Mediterranean Sea during the PROSOPE cruise in summer 1999 (Garczarek *et al.*, 2007), which came to the conclusion that HLI was by far the dominant ecotype in the Mediterranean Sea, at least for the region covered by this cruise that encompassed most of the Western basin and part of the Eastern basin. In contrast, HLII is known to be the dominant HL ecotype in the Red Sea and Arabian Sea (Fuller *et al.*, 2005, 2006). More generally, it was shown earlier that HLI and HLII occupy different latitudinal zones, with HLII constituting the majority of the *Prochlorococcus* population in the upper layer between ~30°S and 35°N and HLI taking over at higher latitudes (Johnson *et al.*, 2006). Our data suggest that the surface waters of the easternmost part of the Mediterranean Sea contain a different *Prochlorococcus* ecotype than do the regions located further west. It is possible that the whole area, including the Cyprus eddy and its vicinity (Tanaka *et al.*, 2007) might have been inoculated by HLII-rich waters that entered the Mediterranean Sea through the Suez Canal.

As for recruitment onto *Gammaproteobacteria* genomes (Figure 3), very few hits were at high identity (>80%). HTCC2080 (NOR5/OM60 clade) recruited a fair amount of hits but only with a 70% identity average. Combined with the BLASTx taxonomic bin scores, this observation points to the fact that currently no marine *Gammaproteobacteria* isolate is well represented in the surface eastern Mediterranean Sea environmental bacteria. This is not surprising, considering that the prevalent *Gammaproteobacteria* members of marine plankton (that is, SAR86) have yet to be cultured.

Clustering of the Mediterranean library by genomic content and metabolic potential

The similarity comparison approach developed by Rusch *et al.* (2007) was used to compare the Mediterranean surface library BAC ends with that of other environments. Fosmid ends obtained from different

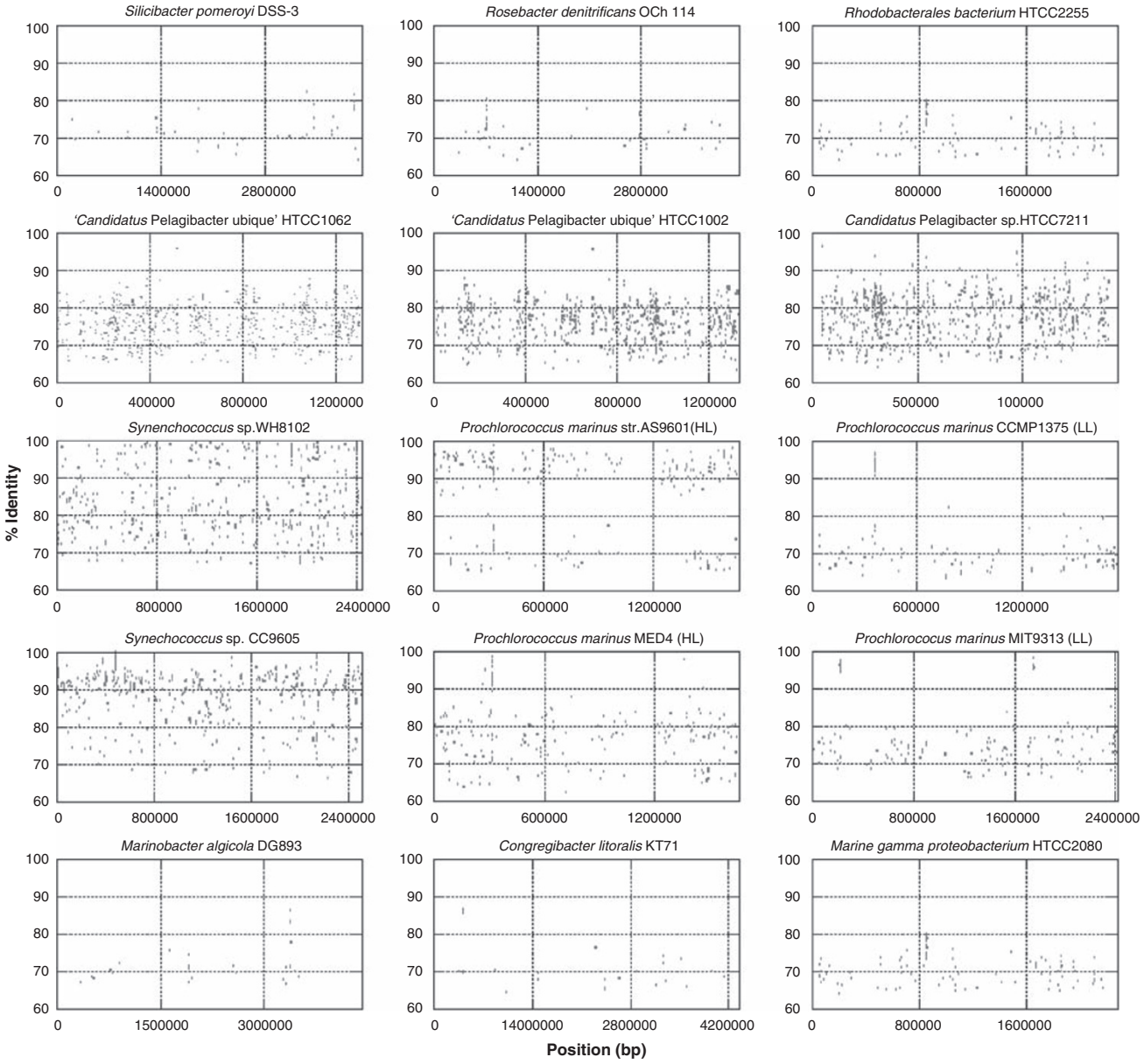

Figure 3 Recruitment of the Mediterranean bacterial artificial chromosome-library end-sequences onto representative genomes of *Alphaproteobacteria*, *Cyanobacteria* and *Gammaproteobacteria*. Red dots indicate BLASTn high-scoring sequence pairs with expectation values of $\leqslant 1 \times 10^{-5}$, >200 bp and covering a minimum 80% of the hit length. The *y* axis indicates the percent identity between the aligned sequence and the genomic sequence. A full colour version of this figure is available at *The ISME Journal* online.

depths of the ALOHA station (DeLong *et al.*, 2006) were compared with ends from a deep Mediterranean fosmid library (3000 m; Martín-Cuadrado *et al.* (2007)) based on protein identities (using tBLASTx). As could be seen in Figure 4a, the Mediterranean surface library clustered with the different ALOHA samples from the photic zone (and also from the 200 m twilight zone) and yet was obviously different from them. Interestingly, when compared on the basis of the COG categories, the Mediterranean surface library clustered with ALOHA surface water, with the sample from ALOHA 10 m being the most similar (Figure 4b), whereas the ALOHA deep stations clustered with the Mediterranean Km3 deep sample. Not surprisingly, these results indicate that the same metabolic needs are shared in similar oceanic layers, independently from their geographic location.

Unique COG categories were much more abundant in the Mediterranean library when compared with the other libraries, including several periplasmic proteins such as F_0F_1-ATP synthase subunits or transporters of categories for substrates such as phosphate, phosphonate, cations, amino acids and sugars (see Supplementary Table S3 for the entire list). When all COGs related to the phosphate or phosphonate metabolism were checked (including those not listed in Supplementary Table S3 as their frequency in the Mediterranean library was lower

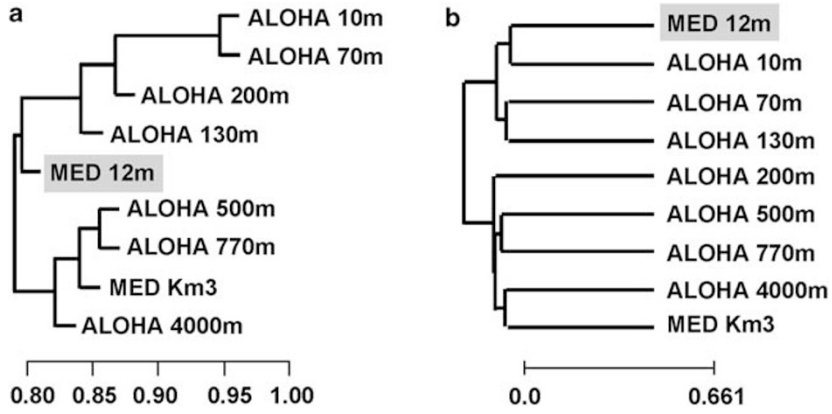

Figure 4 Hierarchical clustering of different environments based on protein similarities or metabolic potential. (**a**) Pairwise similarities of samples based on tBLASTx comparisons (80% identity). (**b**) The 1000 most variable Cluster of Orthologous Group (COG) categories from nine marine environments were normalized (Sharon *et al.*, 2009), standardized and clustered using the Expander statistical package. Raw or normalized COG assignments to the different environments are found in Supplementary Tables S1 and S2, respectively.

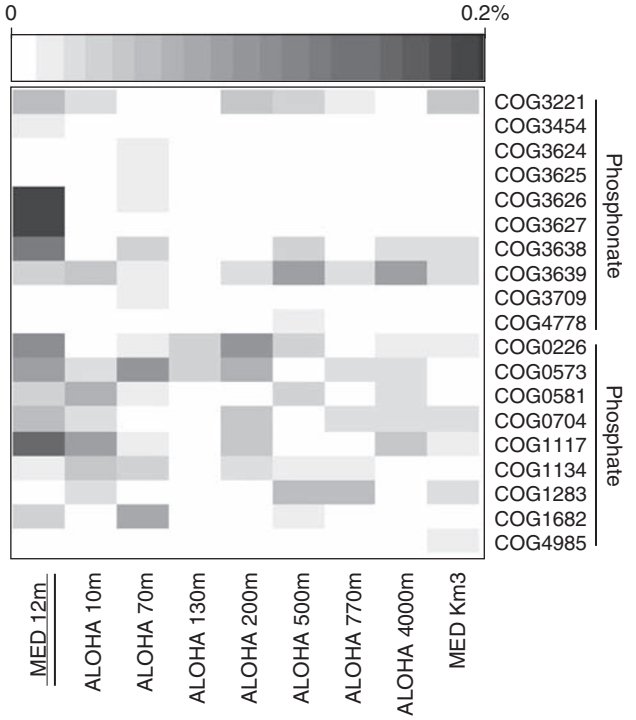

Figure 5 Frequency of different phosphate (lower part of figure) and phosphonate (upper part of figure) metabolism-related Cluster of Orthologous Group (COGs) in the eastern Mediterranean Sea plus eight different marine environments are presented in gray color code gradient. The intensity of the gray shading is proportional to the percentage of each COG category from all COG hits at a specific environment. Raw numbers can be found in Supplementary Table S4.

than 30%), a clearly higher frequency was observed in the MED 12 m library (Figure 5). As the Mediterranean Sea is the largest body of water in the world in which the primary productivity is phosphorus-limited (Krom *et al.*, 1991), bacteria or *Cyanobacteria* possessing the capacity to compete on rare phosphorus sources (Martiny *et al.*, 2006, 2009) are expected to be favored. Interestingly, most BAC-end sequences with similarity to phosphate/phosphonate metabolism COGs were affiliated to different *Alphaproteobacteria* groups, including SAR11. This agrees with a recent report that observes a disproportionately large enrichment in SAR11 periplasmic substrate-binding proteins for phosphate and phosphonate in a metaproteomic study of Sargasso Sea, another phosphorus-limited region (Sowell *et al.*, 2009).

Currently, only few marine BAC or fosmid libraries have been analyzed using random end sequencing, limiting the possible comparisons between different marine environments with regard to the functional diversity of their mostly uncultured microbial populations. However, as more extensive sequencing efforts are now being applied, both to field metagenomes or to cultured strains more representative of natural populations, thanks to the development of high-throughput culturing techniques (see for example, Stingl *et al.* (2007)), our ability to predict the metabolic potential of different environments will increase considerably and allow us to substantially improve our understanding of the functions of naturally occurring microbial communities.

Acknowledgements

This work was partially supported by grants from the state of Lower Saxony and the Volkswagen Niedersachsen Foundation (OB) and grant OCE-0550547 from the US-NSF to MTS and grant OCE-0550547 from the US-NSF to MTS.

References

Alonso-Sáez L, Balagué V, Sà EL, Sánchez O, González JM, Pinhassi J *et al.* (2007). Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* **60**: 98–112.

Béjà O. (2004). To BAC or not to BAC: marine ecogenomics. *Curr Opin Biotech* **15**: 187–190.

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

Cho J-C, Stapels MD, Morris RM, Vergin KL, Schwalbach MS, Givan SA *et al.* (2007). Polyphyletic photosynthetic reaction center genes in oligotrophic marine *Gammaproteobacteria*. *Environ Microbiol* **9**: 1456–1463.

Cho JC, Giovannoni SJ. (2004). Cultivation and growth characteristics of a diverse group of oligotrophic marine *Gammaproteobacteria*. *Appl Environ Microbiol* **70**: 432–440.

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.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.

DeLong EF. (2005). Microbial community genomics in the ocean. *Nat Rev Microbiol* **3**: 459–469.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

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

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP *et al.* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.

Fuchs BM, Spring S, Teeling H, Quast C, Wulf J, Schattenhofer M *et al.* (2007). Characterization of the first marine *gammaproteobacterium* capable of aerobic anoxygenic photosynthesis. *Proc Natl Acad Sci USA* **104**: 2891–2896.

Fuller NJ, Tarran GA, Yallop M, Orcutt KM, Scanlan DJ. (2006). Molecular analysis of picocyanobacterial community structure along an Arabian Sea transect reveals distinct spatial separation of lineages. *Limnol Oceanogr* **51**: 2515–2526.

Fuller NJ, Marie D, Partensky F, Vaulot D, Post AF, Scanlan DJ. (2003). Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69**: 2430–2443.

Fuller NJ, West NJ, Marie D, Yallop M, Rivlin T, Post AF *et al.* (2005). Dynamics of community structure and phosphate status of picocyanobacterial populations in the Gulf of Aqaba, Red Sea. *Limnol Oceanogr* **50**: 363–375.

Garczarek L, Dufresne A, Rousvoal 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 Microbiol Ecol* **60**: 189–206.

Hugenholtz P, Kyrpides NC. (2009). Genomics update: a changing of the guard. *Environ Microbiol* **11**: 551–553.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.

Kan J, Evans SE, Chen F, Suzuki MT. (2008). Novel estuarine bacterioplankton in rRNA operon libraries from the Chesapeake Bay. *Aquat Microbial Ecol* **51**: 55–66.

Kettler G, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**: e231.

Kim UJ, Birren BW, Slepak T, Mancino V, Boysen C, Kang HL *et al.* (1996). Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34**: 213–218.

Kogure K, Simidu U, Taga N. (1979). A tentative direct microscopic method for counting living marine bacteria. *Can J Microbiol* **25**: 415–420.

Krom MD, Brenner S, Kress N, Gordon LI. (1991). Phosphorous limitation of primary productivity in the Eastern Mediterranean Sea. *Limnol Oceanogr* **36**: 424–432.

Kyrpides NC. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol* **27**: 627–632.

Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM *et al.* (2008). The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res* **36**(Database issue): D528–D533.

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 habitat. *PLoS ONE* **2**: e914.

Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.

Martiny AC, Huang Y, Li W. (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340–1347.

Mazard S. (2007). From ecology to genomics: phosphorus acquisition and regulatory mechanisms in marine cyanobacteria mediate niche adaptation. PhD Thesis, Department of Biological Sciences, University of Warwick, Coventry, p 273.

Moeseneder MM, Winter C, Herndl GJ. (2001a). Horizontal and vertical complexity of attached and free-living bacteria of the Eastern Mediterranean Sea, determined by 16S rDNA and 16S rRNA fingerprints. *Limnol Oceanogr* **46**: 95–107.

Moeseneder MM, Arrieta JM, Herndl GJ. (2005). A comparison of DNA- and RNA-based clone libraries from the same marine bacterioplankton community. *FEMS Microbiol Ecol* **51**: 341–352.

Moeseneder MM, Winter C, Arrieta JM, Herndl GJ. (2001b). Terminal-restriction fragment length polymorphism (T-RFLP) screening of a marine archaeal clone library to determine the different phylotypes. *J Microbiol Meth* **44**: 159–172.

Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Nesbø CL, Boucher Y, Dlutek M, Doolittle WF. (2005). Lateral gene transfer and phylogenetic assignment of environmental fosmid clones. *Environ Microbiol* **7**: 2011–2026.

Pedros-Alio C. (2006). Genomics and marine microbial ecology. *Int Microbiol* **9**: 191–197.

Pham VD, Konstantinidis KT, Palden T, DeLong EF. (2008). Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ Microbiol* **10**: 2313–2330.

Pommier T, Canback B, Riemann L, Bostrom KH, Simu K, Lundberg P *et al.* (2007). Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**: 867–880.

Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.

Rappé MS, Kemp PF, Giovannoni SJ. (1997). Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hatteras, North Carolina. *Limnol Oceanogr* **42**: 811–826.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180–1191.

Rusch DB, Halpern AL, Heidelberg KB, Sutton G, Williamson SJ, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: I, the northwest Atlantic through the eastern tropical Pacific. *PLoS Biol* **5**: e77.

Sabehi G, Béjà O. (2007). BAC libraries construction from marine microbial assemblages. In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds). *In Molecular Microbial Ecology Manual*. Kluwer Academic Publishers: Dordrecht, The Netherlands, pp 1863–1879.

Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, Isaacson T *et al.* (2005). New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol* **3**: e173.

Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R *et al.* (2005). EXPANDER-an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**: 232.

Sharon I, Pati A, Markowitz VM, Pinter RY. (2009). A statistical framework for the functional analysis of metagenomes. In: Batzoglou S (ed). *the 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. Springer Berlin/ Heidelberg: Tucson, AZ, USA, pp 496–511.

Sikorski J, Nevo E. (2005). Adaptation and incipient sympatric speciation of *Bacillus simplex* under microclimatic contrast at 'Evolution Canyons' I and II, Israel. *Proc Natl Acad Sci USA* **102**: 15924–15929.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF *et al.* (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3**: 93–105.

Stingl U, Tripp HJ, Giovannoni SJ. (2007). Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J* **1**: 361–371.

Suzuki MT, Béjà O, Taylor LT, DeLong EF. (2001). Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol* **3**: 323–331.

Suzuki MT, Preston CM, Béjà O, de la Torre RJ, Steward GF, DeLong EF. (2004). Quantitative phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microbial Ecol* **48**: 473–488.

Tanaka T, Zohary T, Krom MD, Lawe CS, Pitta P, Psarra S *et al.* (2007). Microbial community structure and function in the Levantine Basin of the eastern Mediterranean. *Deep Sea Res I* **54**: 1721–1743.

Thingstad TF, Krom MD, Mantoura RF, Flaten GA, Groom S, Herut B *et al.* (2005). Nature of phosphorus limitation in the ultraoligotrophic eastern Mediterranean. *Science* **309**: 1068–1071.

Toledo G, Palenik B. (2003). A *Synechococcus* serotype is found preferentially in surface marine waters. *Limnol Oceanogr* **48**: 1744–1755.

West NJ, Scanlan DJ. (1999). Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65**: 2585–2591.

West NJ, Schonhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF *et al.* (2001). Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147**: 1731–1744.

Yan S, Fuchs BM, Lenk S, Harder J, Wulf J, Jiao NZ *et al.* (2009). Biogeography and phylogeny of the NOR5/ OM60 clade of *Gammaproteobacteria*. *Syst Appl Microbiol* **32**: 124–139.

Zwirglmaier K, Heywood JL, Chamberlain K, Woodward MS, Zubkov MV, Scanlan D. (2007). Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. *Environ microbiol* **9**: 1278–1290.

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