The ISME Journal (2013) 7, 256–268

www.nature.com/ismej

Open

ORIGINAL ARTICLE

Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean

Elke Allers[1,4], Jody J Wright[2,4], Kishori M Konwar[2], Charles G Howes[2], Erica Beneze[1], Steven J Hallam[2,3] and Matthew B Sullivan[1]

[1]*Department of Ecology and Evolutionary Biology, University of Arizona, AZ, USA;* [2]*Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada and* [3]*Graduate Program in Bioinformatics, University of British Columbia, Vancouver, British Columbia, Canada*

Marine Group A (MGA) is a candidate phylum of *Bacteria* that is ubiquitous and abundant in the ocean. Despite being prevalent, the structural and functional properties of MGA populations remain poorly constrained. Here, we quantified MGA diversity and population structure in relation to nutrients and O_2 concentrations in the oxygen minimum zone (OMZ) of the Northeast subarctic Pacific Ocean using a combination of catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) and 16S small subunit ribosomal RNA (16S rRNA) gene sequencing (clone libraries and 454-pyrotags). Estimates of MGA abundance as a proportion of total bacteria were similar across all three methods although estimates based on CARD-FISH were consistently lower in the OMZ (5.6% ± 1.9%) than estimates based on 16S rRNA gene clone libraries (11.0% ± 3.9%) or pyrotags (9.9% ± 1.8%). Five previously defined MGA subgroups were recovered in 16S rRNA gene clone libraries and five novel subgroups were defined (HF770D10, P262000D03, P41300E03, P262000N21 and A714018). Rarefaction analysis of pyrotag data indicated that the ultimate richness of MGA was very nearly sampled. Spearman's rank analysis of MGA abundances by CARD-FISH and O_2 concentrations resulted in significant correlation. Analyzed in more detail by 16S rRNA pyrotag sequencing, MGA operational taxonomic units affiliated with subgroups Arctic95A-2 and A714018 comprised 0.3–2.4% of total bacterial sequences and displayed strong correlations with decreasing O_2 concentration. This study is the first comprehensive description of MGA diversity using complementary techniques. These results provide a phylogenetic framework for interpreting future studies on ecotype selection among MGA subgroups, and suggest a potentially important role for MGA in the ecology and biogeochemistry of OMZs.

The ISME Journal (2013) **7,** 256–268; doi:10.1038/ismej.2012.108; published online 15 November 2012
Subject Category: microbial population and community ecology
Keywords: marine; pyrosequencing; bacterial diversity; candidate phylum; oxygen minimum zone; Marine Group A

Introduction

Over the last 25 years, cultivation-independent small subunit ribosomal rRNA (16S rRNA) gene surveys of naturally occurring microbial communities have uncovered a plethora of previously anonymous microbial diversity (Rappé and Giovannoni, 2003; Sogin *et al.*, 2006; Pace, 2009). Although a subset of this diversity can be linked directly to defined phylogenetic groups with cultivated representatives, an increasing number of unaffiliated groups known as candidate phyla or divisions have emerged. Currently, 45 candidate phyla are recognized in public databases on the basis of 16S rRNA gene sequence information, although it is likely that there are many more phyla that have yet to be formally recognized (Rappé and Giovannoni, 2003; McDonald *et al.*, 2011). One of the most prevalent candidate phyla identified in studies of marine microbial diversity is the bacterial Marine Group A (MGA; Fuhrman *et al.*, 1993; Gordon and Giovannoni, 1996; Fuchs *et al.*, 2005; DeLong *et al.*, 2006; Stevens and Ulloa, 2008; Schattenhofer *et al.*, 2009). The first representatives of MGA were described and named as 'Marine Group A' (Fuhrman *et al.*, 1993; Fuhrman and Davis, 1997) or the 'SAR406 gene lineage' (Gordon and Giovannoni, 1996) based on 16S rRNA gene sequence information collected from Atlantic and Pacific Ocean waters. Contemporary

Correspondence: SJ Hallam, Department of Microbiology and Immunology, University of British Columbia, 2552-2350 Health Sciences Mall, Vancouver, British Columbia, V6T 1Z3 Canada.
E-mail: shallam@interchange.ubc.ca
or MB Sullivan, Department of Ecology and Evolutionary Biology, University of Arizona, PO Box 210088, Tucson, Arizona 85721, USA.
E-mail: mbsulli@email.arizona.edu
[4]These authors contributed equally to this work.
Received 1 May 2012; revised 30 July 2012; accepted 9 August 2012; published online 15 November 2012

phylogenetic analyses indicate that MGA is most closely related to the phylum *Calditrix,* named after a genus of anaerobic, mixotrophic, thermophiles obtained from a hydrothermal vent chimney in the Mid-Atlantic Ridge (Miroshnichenko *et al*., 2003; Rappé and Giovannoni, 2003).

MGA are most prevalent below the photic zone in stratified waters with distinct halo or oxyclines. Indeed, in surveys of oxygen minimum zones (OMZs) and permanent or seasonally stratified anoxic basins, 16S rRNA gene sequences affiliated with MGA are well represented in clone libraries (Madrid *et al*., 2001; Fuchs *et al*., 2005; Stevens and Ulloa, 2008; Zaikova *et al*., 2010; Wright *et al*., 2012). In these systems, O_2 serves as a key organizing principle for microbial community structure and function, defining specific metabolic niches and biogeochemical potentials across the oxycline (Wright *et al*., 2012). MGA are particularly diverse and abundant within the OMZ of the Northeast subarctic Pacific Ocean (NESAP; Wright *et al*., 2012). The distinct and well-studied coastal to open-ocean gradients of biological production, nutrients and O_2 existing within the OMZ of the NESAP make it an ideal natural laboratory in which to explore ecological and biogeochemical roles of MGA in the ocean. Here, we use a combination of catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH), 16S rRNA gene clone libraries and pyrotag sequencing to quantify MGA abundance and diversity along the Line P oceanographic transect of the NESAP (Pena and Bograd, 2007). We then apply statistical analyses to explore the hypothesis of O_2 and other environmental factors as drivers of habitat selection for different MGA subgroups in the water column.

Materials and methods

Sample collection and processing

Sampling was conducted via multiple hydrocasts using a conductivity, temperature, depth rosette water sampler aboard the *CCGS John P Tully* during Line P cruise 2009-09 in the NESAP in June 2009 (major stations: P4 (48°39.0°N, 126°4.0°W)—7 June, P12 (48°58.2°N, 130°40.0°W)—9 June and P26 (50°N, 145°W)—14 June). At these three stations, large volume (20 l) samples for DNA isolation were collected from the surface (10 m), whereas 120 l samples were taken from three depths spanning the OMZ core and upper and deep oxyclines (500 m, 1000 m and 1300 m at station P4 and 500 m, 1000 m and 2000 m at stations P12 and P26). Sample collection and filtration protocols can be viewed as visualized experiments at http://www.jove.com/video/1159/ (Zaikova *et al*., 2009) and http://www.jove.com/video/1161/ (Walsh *et al*., 2009a), respectively.

For small-volume sampling (for CARD-FISH), the water from Niskin bottles was transferred into pre-rinsed 1-l plastic bottles, filtered through a 10-μm nylon mesh filter and processed immediately. The conductivity, temperature, depth-mounted O_2 probe (Model SBE 43, Sea-Bird Electronics, Bellevue, WA USA) reported O_2 concentrations in $\mu mol\,kg^{-1}$. Nutrient samples were collected in plastic tubes and analyzed at sea (stored at 4 °C and in the dark before analysis) using an Astoria Analyzer (Astoria-Pacific, Clackamas, OR, USA) as described by Barwell-Clarke and Whitney (1996).

Chlorophyll a

Chlorophyll *a* (Chl*a*) was measured *in situ* with a Seapoint chlorophyll fluorometer (Seapoint Sensors, Exeter, NH, USA) and ground-truthed with 109 selected reference samples collected on 47 mm GF/F filters (Whatman International, Maidstone, UK) for Chl*a* extraction (Holm-Hansen *et al*., 1965). The linear regression between reference sample fluorescence and Chl*a* data were used to transform depth corrected fluorescence units to Chl*a* (Cuttelod and Herve, 2010; $R^2 = 0.90$; data not shown).

Enumeration of cells by flow cytometry

Cells were enumerated by flow cytometry using samples fixed with formaldehyde (final concentration of 4% wt/vol) and stored at 4 °C until analysis using SYBR Green I (Invitrogen, Carlsbad, CA, USA) on a FACS LSRII (Becton Dickonson, Franklin Lakes, NJ, USA; Zaikova *et al*. 2010).

Catalyzed reporter deposition fluorescence in situ hybridization

Pre-filtered (10 μm) seawater samples were fixed with formaldehyde (16%, Polysciences, Warrington, PA, USA) at a final concentration of 1–2% at 4 °C for 12–24 h. Subsamples were filtered onto 47 mm 0.2 μm membrane filters (GTTP, Millipore, Billerica, MA, USA) and rinsed with Milli-Q water. Filters were left to air dry and then stored at −80 °C until analysis by CARD-FISH as described by Pernthaler *et al*. (2004). In brief, cells were fixed to the filter membrane by agarose embedding. Endogenous peroxidases were inactivated by HCl treatment, cells were permeabilized by lysozyme (for probes EUBI-III (Amann *et al*., 1990; Daims *et al*., 1999), NON338 (Wallner *et al*., 1993), SAR406-97 (Fuchs *et al*., 2005)) or a combination of lysozyme and achromopeptidase or HCl (tested for optimization only). For hybridization, horseradish peroxidase-labeled probes EUBI-III and NON338, and SAR406-97 were added to hybridization buffers containing 35% and 40% formamide (Fisher, Pittsburg, PA, USA), respectively. Hybridizations were performed at 46 °C and followed by washing steps to remove unspecifically bound probe. During the CARD step, the dye Alexa Fluor488 (Invitrogen, Molecular Probes, Carlsbad, CA) was combined with the remaining substrate mix

at 1:300. The fraction of FISH-stained bacteria was quantified microscopically at $\times 1000$ magnification in at least 1000 4′6-diamidino-2-phenyl indole-stained cells in 10 or more fields of vision per sample using an AxioImager (Zeiss, Jena, Germany).

Environmental DNA extraction for 16S rRNA gene clone library construction

DNA was extracted from sterivex filters as described in Zaikova *et al.* (2010) and DeLong *et al.* (2006). The DNA extraction protocol can be viewed as a visualized experiment at http://www.jove.com/video/1352/ (Wright *et al.*, 2009).

PCR amplification of 16S rRNA gene, clone library construction and sequencing

A total of 12 DNA extracts from large volume samples collected from four depths at stations P4, P12 and P26 in February 2009 (using the same sampling plan and protocols described above) were amplified using small subunit ribosomal DNA (16S rRNA gene) primers targeting the bacterial domain: B27F (5′-AGAGTTTGATCCTGGCTCAG-3′) and U1492R (5′-GGTTACCTTATGTACGACTT-3′) under the following PCR conditions: 3 min at 94 °C followed by 35 cycles of 94 °C for 40 s, 55 °C for 1.5 min, 72 °C for 2 min and a final extension of 10 min at 72 °C. Each 50 μl reaction contained 1 μl of DNA, 1 μl each 10 mM forward and reverse primer, 2.5 U Taq (Qiagen, Germantown, MD, USA), 5 μl 10 mM deoxynucleotides and 41.5 μl 1X Qiagen PCR Buffer. 16S rRNA gene amplicons were purified, transformed and cloned as described previously (Zaikova *et al.*, 2010-3′) with the following modifications: one 384-well plate per depth interval was picked and sent for Sanger sequencing at the Michael Smith Genome Sciences Centre (Vancouver, British Columbia, Canada). Sequence data were collected on an AB 3730xls (Applied Biosystems, Carslbad, CA, USA). Plasmids were sequenced bidirectionally with M13F (5′-GTAAAACGACGGCCAG-3′) and M13R (5′-CAGGAAACAGCTATGAC-3′) primers. Bidirectional sequence reads were assembled using Sequencher v4.8 (Gene Codes Corporation, Ann Arbor, MI, USA) and manually edited for base-calling errors. The resulting data sets were checked for chimeras with the open source application Bellerophon (Huber *et al.*, 2004; using default settings) and 745 chimeric sequences were removed.

Phylogenetic analysis and tree construction using MGA 16S rRNA gene sequences

A total of 3164 non-chimeric 16S rRNA gene sequences were imported into the ARB software package (release 106; Ludwig *et al.*, 2004). Sequences were added to the full-length SILVA database (http://www.arb-silva.de; Pruesse *et al.*, 2007), aligned to the closest relative and added to an existing tree of sequences from the ARB database by using the ARB parsimony tool (using default parameters).

A maximum likelihood phylogenetic tree of MGA 16S rRNA gene sequences exported from ARB was inferred by PHYML (Guindon *et al.*, 2005) using an HKY + 4G + I model of nucleotide evolution where the parameter of the gamma distribution, the proportion of invariable sites and the transition/transversion ratio were estimated for each data set. The confidence of each node was determined by assembling a consensus tree of 100 bootstrap replicates. Bacterial 16S rRNA gene sequences (including 170 previously published sequences) generated from the Line P transect in June 2008 (station P4 1000 m; Walsh *et al.*, 2009b) were also placed in taxonomic hierarchy for downstream analysis using the NAST aligner (DeSantis *et al.*, 2006b) and blast using default parameters against the 2008 Greengenes database (DeSantis *et al.*, 2006a), and 290 sequences were identified as belonging to MGA. These 290 sequences were clustered at 97% identity using mothur (v.1.19.0; Schloss *et al.*, 2009). Representative sequences from each of these clusters were identified using the get.oturep command in mothur and were included in the phylogenetic tree.

PCR amplification of 16S rRNA gene for pyrotag sequencing

To more directly compare the quantitative distribution of MGA in relation to CARD-FISH counts, the V6–V8 region of 16S rRNA was amplified from June 2009 DNA samples using primers 926F (5′-cctatcccctgtgtgccttggcagtctcag AAACTYAAAKGAA TTGRCGG-3′) and 1392R (5′-ccatctcatccctgcgtgtctccgactcag-<**XXXXX**>-ACGGGCGGTGTGTRC-3′). Primer sequences were modified by the addition of 454A or B adapter sequences (lower case). In addition, the reverse primer included a 5-bp barcode designated <**XXXXX**> for multiplexing of samples during sequencing. Twenty-microlitre PCR reactions were performed in duplicate and pooled to minimize PCR bias using 0.4 μl Advantage GC 2 Polymerase Mix (Advantage-2 GC PCR Kit, Clonetech, Mountainview, CA, USA), 4 μl 5X GC PCR buffer, 2 μl 5 M GC Melt Solution, 0.4 μl 10 mM dNTP mix (MBI Fermentas, Glen Burnie, MA, USA), 1.0 μl of each 25 nM primer and 10 ng sample DNA. The thermal cycler protocol was 95 °C for 3 min, 25 cycles of 95 °C for 30 s, 50 °C for 45 s, and 68 °C for 90 s and a final 10-min extension at 68 °C. PCR amplicons were purified using SPRI Beads and quantified using a Qubit fluorometer (Invitrogen). Samples were diluted to 10 ng μl^{-1} and mixed in equal concentrations. Emulsion PCR and sequencing of the PCR amplicons were performed at the Department of Energy Joint Genome Institute (Walnut Creek, CA, USA) following the Roche 454 GS FLX Titanium (454 Life Sciences, Branford, CT,

USA) technology according to the manufacturer's instructions.

Processing of pyrotag sequences
A total of 219 610 pyrotag sequences were analyzed using the Quantitative Insights Into Microbial Ecology (QIIME) software package (Caporaso *et al.*, 2010). Reads with length <200 bases, ambiguous bases and homopolymer runs were removed before chimera detection. Chimeras were detected using the chimera slayer provided in the QIIME software package and removed before taxonomic analysis. A total of 212 611 non-chimeric sequences were phylogenetically identified in QIIME using a BLAST-based assignment method and clustered at 97% identity against the Greengenes taxonomic database (DeSantis *et al.*, 2006a). Singleton operational taxonomic units (OTUs; represented by one read) were omitted from downstream analyses, as recommended by Kunin *et al.* (2010), Tedersoo *et al.* (2010) and Gihring *et al.* (2012), leaving 183 212 sequences for downstream analysis.

Clustering of pyrotags to 16S rRNA gene clone library sequence clusters
To resolve patterns of distribution among MGA clusters as a function of geographic location in the NESAP, pyrotag sequences were recruited to MGA 16S rRNA gene clone library sequence clusters using a 97% identity cutoff in mothur. Blastn was used to query 183 212 pyrotags against a database containing 290 16S rRNA gene clone library sequences assigned to MGA based on Greengenes taxonomy. Only hits with a perfect match across the full length of a query sequence were retrieved, and the number of pyrotags mapping to all sequences in each cluster was summed. If a pyrotag mapped to >1 cluster, its relative contribution to each cluster was calculated by dividing by the number of clusters it mapped to and assigning the relevant fraction to each cluster. The number of pyrotags mapping to each cluster was normalized to the total number of bacterial tags in each sample (Table 1) and visualized as a bubble plot using bubble.pl, available for download at http://www.cmde.science.ubc.ca/hallam/bubble.php. A rarefaction curve for full-length MGA 16S rRNA sequences and MGA pyrotag sequences was calculated and plotted using QIIME (Caporaso *et al.*, 2010).

Estimating probe SAR406-97 detection efficiency
To test the predicted maximum binding efficiency of probe SAR406-97 (Fuchs *et al.*, 2005; 5′-CACC CGTTCGCCAGTTTA) against MGA 16S rRNA gene clone library sequences from the NESAP, blastn (*E*-value = 1000, word_size = 7) was used to query the probe sequence against the 290 16S rRNA gene clone library sequences assigned to MGA based on Greengenes taxonomy and collect all local alignments with similarity to the probe sequence. Probe efficiency was described using the percentage of MGA sequences that contained local alignments to the probe across a range of *E*-value scores for each cluster.

Results

Physicochemical characteristics of the NESAP
Our study site, Line P, is a 1425-km survey line of the NESAP, originating in Saanich Inlet, British Columbia, Canada (SI; 48°N, 123°W), and terminating at Ocean Station Papa (also known as station P26; 50°N, 145°W), on the southeast edge of the Alaskan Gyre (Pena and Bograd, 2007; Pena and Varela, 2007;

Table 1 Chemical and biological parameters at Line P stations P4, P12 and P26 in June 2009

Station	*Depth (m)*	*Microbial cell abundance by FCM (cells per ml)*	*MGA*[a] *(% total DAPI cell number)*	*No. of bacterial 16S rRNA clones*[b]	*MGA clones (%16S rRNA clone library)*	*No. of bacterial pyrotags*	*MGA pyrotags*[c] *(% bacterial pyrotags in library)*	*Oxygen ($\mu mol\,kg^{-1}$)*	*Nitrate*[d] *($\mu mol\,l^{-1}$)*
P4	10	1.25E+05	1.3	281	0.0	11 178	0.1	308.0	0.0
	500	1.23E+04	7.8	287	8.7	14 619	7.8	23.7	42.6
	1000	2.22E+04	6.7	276	9.8	6251	11.6	8.6	45.3
	1300	1.86E+04	3.5	239	10.5	15 284	11.5	15.9	45.8
P12	10	1.21E+05	0.7	248	1.6	14 189	0.1	296.2	6.3
	500	1.66E+04	3.8	249	4.0	11 759	10.3	37.0	42.8
	1000	7.89E+03	6.7	184	13.6	14 839	8.1	9.0	46.3
	2000	7.36E+03	5.5	256	13.7	7391	9.3	59.3	44.1
P26	10	1.41E+05	0.4	242	0.4	11 723	0.4	301.4	10.8
	500	1.47E+04	3.3	287	7.7	7648	8.3	35.0	43.6
	1000	1.72E+04	8.2	293	16.4	7901	9.6	14.3	45.6
	2000	8.78E+03	4.5	322	14.3	16 090	13.0	56.5	44.4

Abbreviations: DAPI, 4′6-diamidino-2-phenyl indole; FCM, flow cytometry; MGA, Marine Group A; 16S rRNA, small subunit ribosomal rRNA.
[a]MGA as detected by probe SAR406-97 as fraction of total cell count of DAPI-stained cells.
[b]16S rRNA gene clone libaries were generated from February 2009 samples.
[c]MGA pyrotags taxonomically identified by comparison with Greengenes database.
[d]Nitrate + nitrite.

Supplementary Figure S1). The NESAP is characterized by strong stratification with a maximum winter mixing depth of 125–150 m (Freeland, 1997; Whitney *et al.*, 1998). As such, the interior regions of the NESAP are insulated from the atmosphere, creating a vast OMZ centered at 1000 m with oxyclines extending from ~400 m to 2000 m containing O_2 concentrations ranging between ~9 μmol kg^{-1} and 60 μmol kg^{-1} (Freeland, 1997; Whitney *et al.*, 2007). The NESAP OMZ (also referred to as the Eastern Subtropical North Pacific OMZ) is the largest and least studied permanent OMZ in the global ocean (Paulmier and Ruiz-Pino, 2009).

Relevant physicochemical data from representative coastal (P4), transition (P12) and open-ocean (P26) stations measured along the Line P transect and related to this study are described below. Salinity gradients ranging from 32.2 PSU to 32.6 PSU at the surface (10 m) and 34.1–34.6 PSU in the ocean's interior generated a stratified water column across the Line P transect (Supplementary Figure S2). Chl*a* was present in the top ~100 m, with deep chlorophyll maxima ranging from 0.5 mg l^{-1} at 41 m depth at P26 to 1.1 mg l^{-1} at 25 m depth at P4 (Figure 1). Average O_2 concentrations were 302 μmol kg^{-1} at the surface, reaching a minimum of 8.6–15 μmol kg^{-1} between 1000 m and 1100 m across the transect (Table 1, Figure 2). The OMZ core (defined as $O_2 < 20$ μM (~19.5 μmol kg^{-1}); Helly and Levin, 2004; Paulmier and Ruiz-Pino, 2009) was 766 ± 73 m thick and centered at 1026 ± 63 m. Nutrient concentrations were higher in the OMZ core and the upper (500 m) and deep (2000 m) oxyclines than at the surface (Table 1, Supplementary Figure S2). In 10 m samples, nitrate and phosphate concentrations were highest at P26 (9.9 μmol l^{-1} and 1.0 μmol l^{-1}, respectively). At 1000 m, nitrate concentration was highest at P26 (47.5 μmol l^{-1}), whereas phosphate concentration was highest at P4 (3.3 μmol l^{-1}). All contextual data are available through the Canadian Department of Fisheries and Oceans (http://www.pac.dfo-mpo.gc.ca/science/oceans/data-donnees/line-p/).

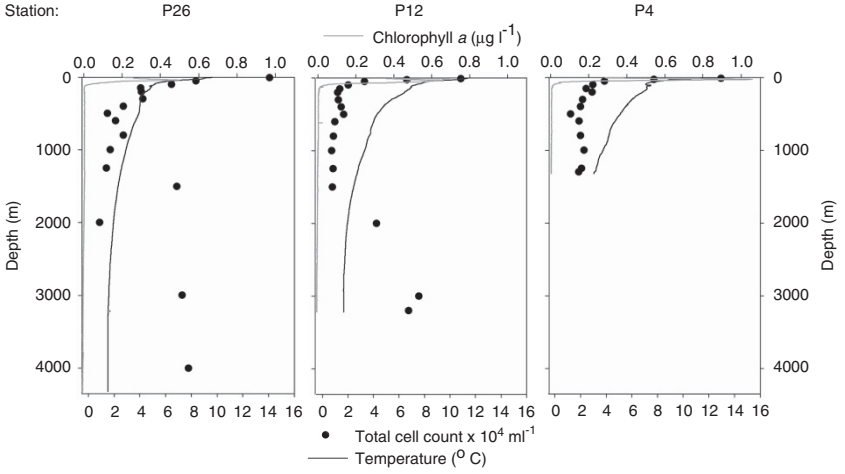

Figure 1 Contextual data for Line P stations P4, P12, and P26 in June 2009. Depicted are Chl*a*, temperature, and total cell counts detected by flow cytometry. A full-colour version of this figure is available at *The ISME Journal* Online.

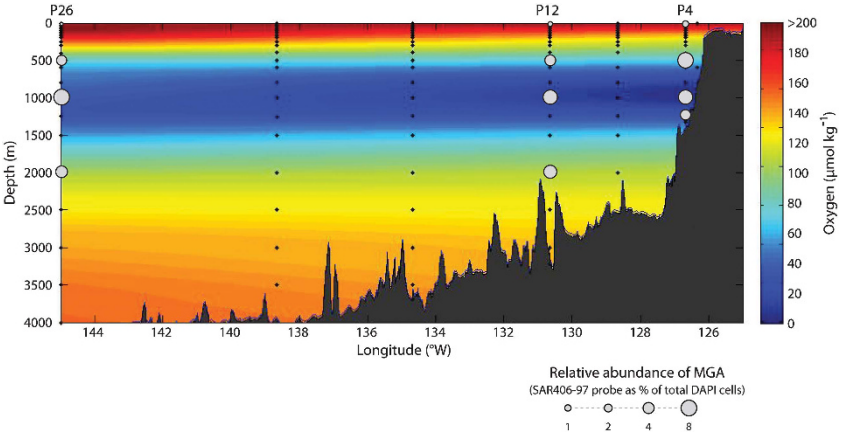

Figure 2 Relative abundance of MGA by CARD-FISH in the NESAP at Line P stations P4, P12 and P26 in June 2009. O_2 concentration is depicted as colored background and MGA abundance is overlaid as gray bubbles.

Microbial cell numbers

Total microbial abundance along the Line P transect was $(1.3 \pm 0.1) \times 10^5\,ml^{-1}$ in surface waters and $(1.39 \pm 0.2) \times 10^4\,ml^{-1}$ in waters >200 m as measured by flow cytometry (Table 1, Figure 1). The overall detection of *Bacteria* by probes EUBI-III ranged from 25.5% ± 7.6% to 79.5% ± 8.6% of total 4′6-diamidino-2-phenyl indole cell counts with higher detection rates in surface samples (Supplementary Table S1). Low EUB detection did not appear to result from poor cell lysis, as comparison of lysozyme vs lysozyme/achromopeptidase treatment (Pernthaler *et al*., 2004) revealed no significant differences (data not shown). Sequence comparison by BLAST analysis suggested that >90% of our full-length bacterial 16S rRNA gene clone library sequences were targeted by EUBI-III probes with an *E*-value of 10^{-4} (corresponding to a blastn result with no mismatches and up to one missing 3′ base; data not shown).

Diversity and population structure of MGA

Relative abundance of MGA cells as detected by probe SAR406-97 was similar at stations P4, P12 and P26, with minima in surface waters and maxima in waters ≥500 m (≤1.3% vs ∼8%, respectively; Figure 2). At stations P12 and P26, MGA abundance peaked in the core of the OMZ (6.7% ± 1.8% and 8.2% ± 1.6%, respectively) with lower values (3.3–5.5%) in the upper and deep oxyclines (Table 1, Figure 2). At station P4, MGA abundance peaked in the upper oxycline (7.8% ± 2.3%) and decreased throughout the OMZ core and deep oxycline. Blastn-based sequence comparisons against our full-length 16S rRNA gene clone library sequences suggested that probe SAR406-97 targeted ∼76% of all MGA sequences (see below) with an *E*-value of 10^{-4} (corresponding to a blastn result with no mismatches and up to one missing 3′ base; Supplementary Tables S2a and b).

A total of 290 MGA 16S rRNA gene sequences were recovered from 3164 bacterial sequences traversing the water column at stations P4, P12 and P26. MGA sequences comprised an average of 0.7% ± 0.84% of 10 m clone libraries and 11.2% ± 3.9% of libraries from O_2-deficient waters (<90 μmol kg^{-1} O_2) with a maximum of 16.4% at P26 1000 m (Table 1). MGA 16S rRNA gene sequences clustered at 97% identity into 121 distinct OTUs, 97 of which contained only singletons (Supplementary Table S2). Representative sequences obtained for each OTU were placed in phylogenetic context with relevant reference sequences (Figure 3). Five previously defined subgroups were recovered (ZA3648c and ZA3312c (Fuchs, unpublished), Arctic96B-7 and Arctic95A-2 (Bano and Hollibaugh, 2002) and SAR406 (Gordon and Giovannoni, 1996)), and five additional subgroups were defined (HF770D10, P262000D03, P41300E03, P262000N21 and A714018). The most abundant OTUs present along the Line P transect comprised between 1% and 4% of at least one clone library and belonged to subgroups Arctic95A-2, HF770D10, SAR406, Arctic96B-7 and ZA3312c (Figure 4, Supplementary Table S2a).

To explore the diversity and population structure of MGA subgroups with increased resolution, we performed 454-pyrotag sequencing (Table 1). Pyrotags affiliated with MGA OTUs were identified using two approaches: (1) recruitment of pyrotags to full-length 16S rRNA gene sequences and (2) direct taxonomic assignment of pyrotags in blast-based queries to identify OTUs not detected in clone libraries.

In the first approach, we recruited all pyrotags to all 16S rRNA gene clone library sequences affiliated with MGA (see Materials and methods). A total of 4403 pyrotags formed identical matches to 78 out of 121 previously defined MGA OTUs (Figure 4). The relative proportion of bacterial pyrotags affiliated with MGA OTUs ranged from ∼0.01% in 10 m samples to a maximum of 5.7% at P4 1000 m. Within O_2-deficient waters, the average proportion of bacterial pyrotags belonging to MGA was 4.4% ± 0.73%. The most abundant MGA OTUs based on pyrotag recruitment were affiliated with Arctic95A-2 (∼2.4%), Arctic96B-7 (0.55%), SAR406 (∼0.45%), HF770D10 (0.55%) and A714018 (0.26%).

In the second approach, all non-singleton pyrotags were queried against the Greengenes database (DeSantis *et al*., 2006a) resulting in the identification of 10 278 sequences affiliated with MGA (Figure 5a). The relative proportion of bacterial pyrotags affiliated with MGA ranged from ∼0.1% in 10 m samples to a maximum of 11.6% at P4 1000 m (Table 1). Within O_2-deficient waters, the average proportion of bacterial pyrotags belonging to MGA was 9.9% ± 1.8%. To identify MGA OTUs unique to pyrotags, we extracted the corresponding V6–V8 region from the 290 16S rRNA gene clone library sequences identified as MGA and clustered these with the subset of pyrotag sequences affiliated with MGA at 97% identity into 566 distinct OTUs, 491 of which were unique to pyrotags (Figure 5b). However, the majority of abundant OTUs (containing >200 sequences) were common between 16S rRNA gene clone libraries and pyrotag data sets (Figure 5c). Of the unique pyrotag OTUs, 249 were non-singleton and contained 4253 pyrotags (40% of MGA pyrotags), with the most abundant OTU containing 1409 sequences (13.3% of MGA pyrotags; Figure 5c). The slope of the rarefaction curve for MGA pyrotags became nearly asymptotic, indicating that the ultimate richness of MGA OTUs was very nearly sampled (Supplementary Figure S3). In contrast, the rarefaction curve for MGA 16S rRNA gene clone library sequences indicated incomplete sampling.

Comparing MGA abundance across methods

To evaluate consistency in estimating MGA abundance using CARD-FISH, 16S rRNA gene clone libraries and pyrotags, Spearman's rank correlation

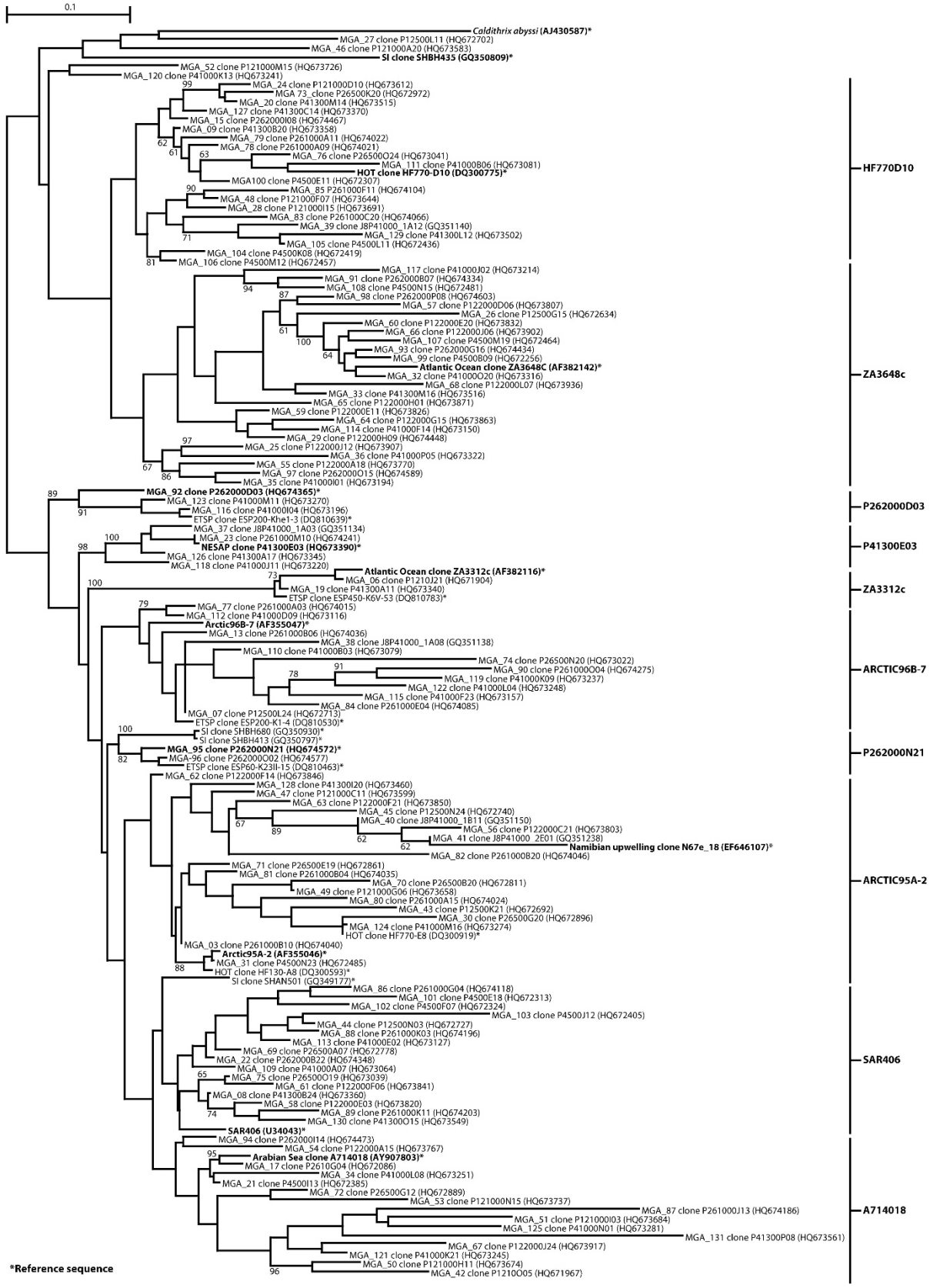

Figure 3 Unrooted phylogenetic tree based on 16S rRNA gene clone sequences showing the phylogenetic affiliation of MGA sequences identified in this study. The tree was inferred using maximum likelihood implemented in PhyML (Guindon *et al.*, 2005). Reference sequences from other environments are marked with an asterisk. The bar represents 10% estimated sequence divergence.

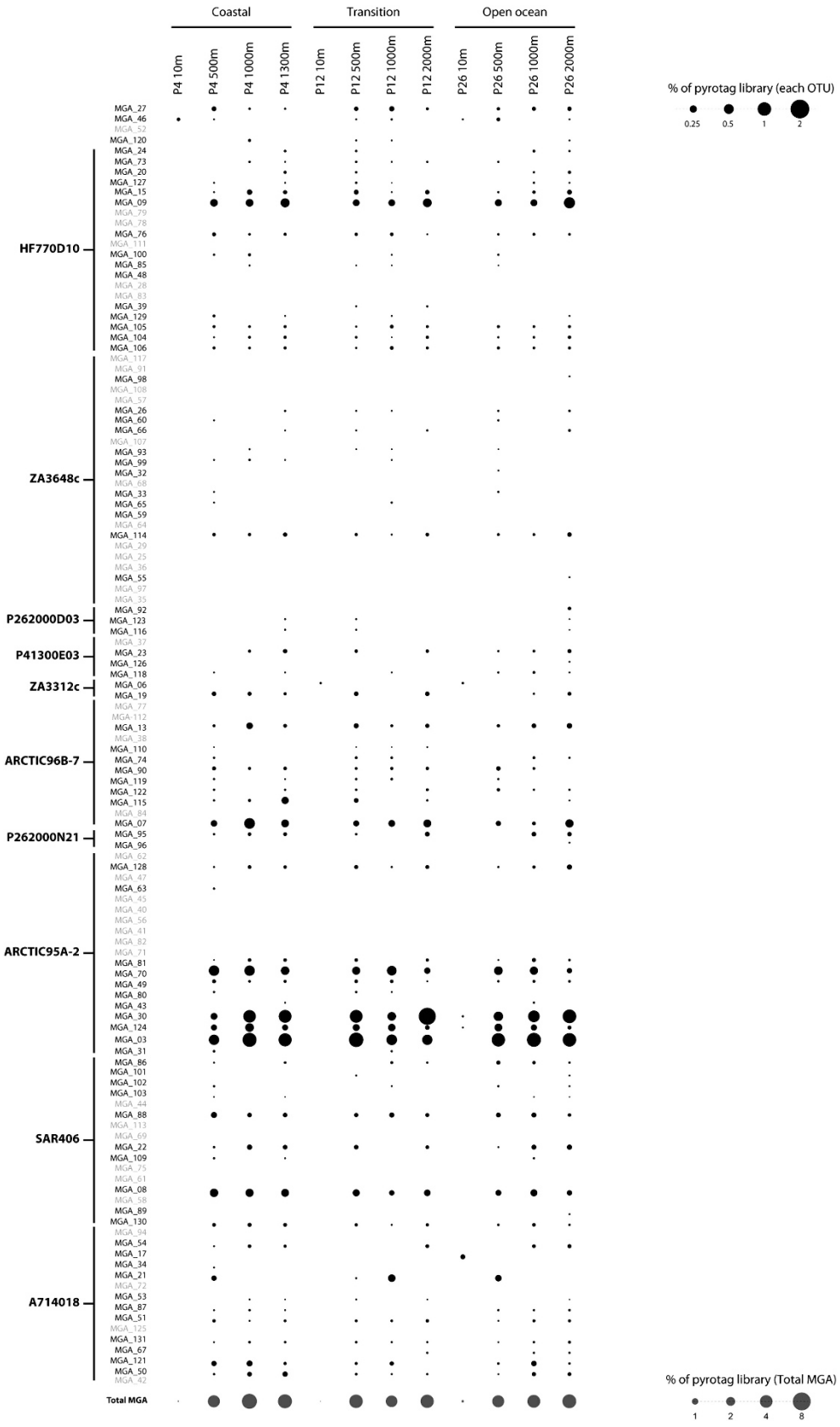

Figure 4 Relative abundance of MGA pyrotags affiliated with full-length MGA 16S rRNA gene clone OTUs recovered from the Northeast subarctic Pacific Ocean. Black circles represent proportion of bacterial pyrotags affiliated with each 16S rRNA OTU in each sample. A full-colour version of this figure is available at *The ISME Journal* Online.

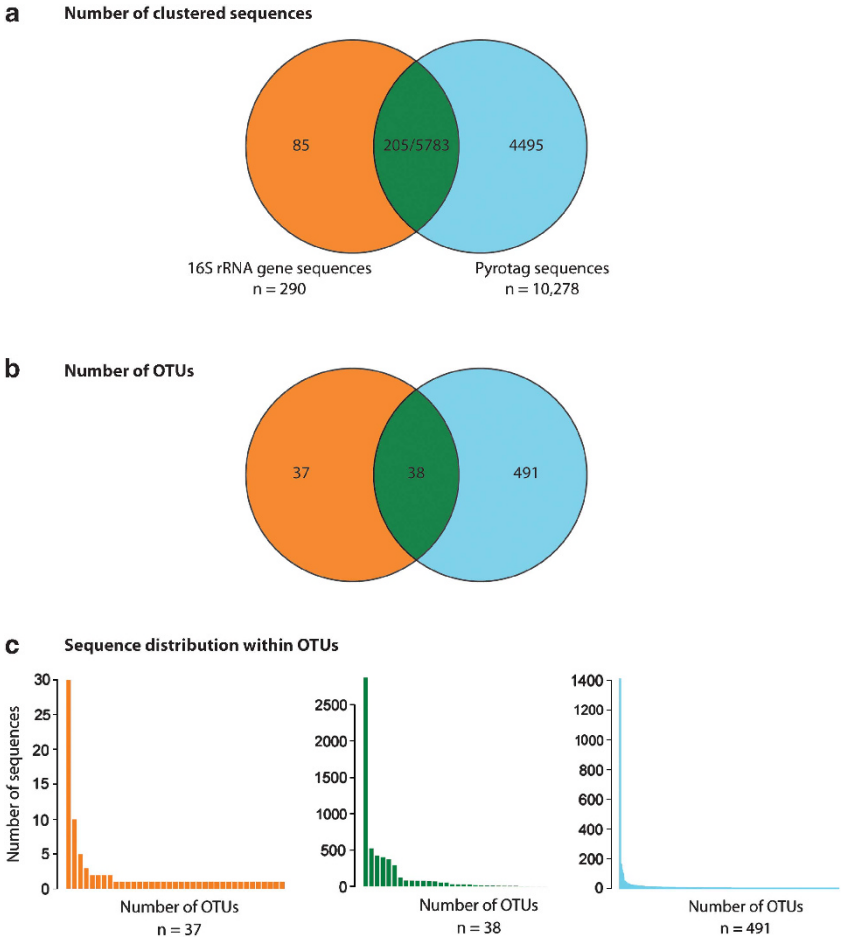

Figure 5 Comparison of V6–V8 region of full-length 16S rRNA gene clone sequences affiliated with MGA, and pyrotags taxonomically identified as MGA by comparison with Greengenes. (**a**) Number of MGA sequences shared between and unique to 16S rRNA gene clone libraries and pyrotags. (**b**) Number of MGA OTUs shared between and unique to 16S rRNA gene clone libraries and pyrotags. (**c**) Sequence distribution within shared and unique MGA OTUs.

Table 2 Spearman's rank correlation coefficients between relative abundance of MGA estimated by CARD-FISH[a] and environmental parameters

Station	n	*Depth*	*Temperature*	*Oxygen*	*Salinity*	*Chla*	*Nitrate*	*Phosphate*	*Silicate*
P4	13	0.396	−0.385	−0.396	0.396	−0.358	0.396	0.429	0.396
P12	5	0.9	−0.9	−0.3	0.9	−0.9	0.4	0.359	0.9
P26	17	0.701*	−0.701*	−0.824**	0.699*	−0.514*	0.865**	0.853**	0.706*
All	35	0.620**	−0.577**	−0.589**	0.621**	−0.553**	0.639**	0.578**	0.623**

Abbreviations: CARD-FISH, catalyzed reporter deposition fluorescence *in situ* hybridization; MGA, Marine Group A.
*$P<$ 0.050; **$P<$ 0.001; n = number of samples.
[a]Using probe SAR406-97.

coefficients (ρ) were determined (Supplementary Figure S4). CARD-FISH abundance estimates were significantly correlated ($P<0.05$) with 16S rRNA gene clone library sequence abundance ($\rho=0.755$) but not with pyrotag sequence abundance ($\rho=0.469$; Supplementary Figures S4a and b). 16S rRNA gene clone library and pyrotag sequence abundance were also significantly correlated ($\rho=0.580$, Supplementary Figure S4c).

To explore potential drivers of MGA habitat selection, we calculated Spearman's rank correlation coefficients between CARD-FISH, 16S rRNA gene clone library, and pyrotag sequence abundance and environmental parameters. When calculated across the entire transect, the abundance of MGA as estimated by CARD-FISH was significantly correlated with decreasing temperature, O_2 and Chl*a*, and increasing nitrate, phosphate and silicate (Table 2).

Table 3 Pyrotag OTUs with statistically significant Spearman's rank correlations (ρ) with oxygen concentration (17 out of 79) in the NESAP

Pyrotag OTU	*Depth* ρ	*Temperature* ρ	*Salinity* ρ	*Oxygen* ρ	*Nitrate* ρ	*Phosphate* ρ	*Silicate* ρ	*Chla* ρ	*Phylogenetic affiliation*
MGA_100	0.000	0.071	0.017	−0.616*	0.166	0.558*	−0.071	0.067	HF770_D10
MGA_105	0.670*	−0.648*	0.634*	−0.648*	0.683*	0.676*	0.648*	−0.627*	HF770_D10
MGA_106	0.670*	−0.648*	0.634*	−0.648*	0.683*	0.676*	0.648*	−0.627*	HF770_D10
MGA_76	0.407	−0.366	0.380	−0.718*	0.549	0.630*	0.366	−0.500	HF770_D10
MGA_99	0.217	−0.058	0.221	−0.761*	0.358	0.707*	0.058	−0.096	ZA3648c
MGA_90	0.330	−0.256	0.359	−0.580*	0.342	0.576*	0.256	−0.313	Arctic96B-7
MGA_07	0.854**	−0.732*	0.810*	−0.599*	0.599*	0.637*	0.732*	−0.718*	Arctic96B-7
MGA_03	0.472	−0.514	0.423	−0.648*	0.620*	0.634*	0.514	−0.486	Arctic95A-2
MGA_49	0.386	−0.345	0.359	−0.683*	0.507	0.595*	0.345	−0.486	Arctic95A-2
MGA_88	0.328	−0.331	0.359	−0.824**	0.613*	0.768*	0.331	−0.331	Arctic95A-2
MGA_124	0.312	−0.326	0.291	−0.827**	0.606*	0.771*	0.326	−0.396	Arctic95A-2
MGA_70	0.342	−0.275	0.359	−0.880**	0.542	0.832**	0.275	−0.289	Arctic95A-2
MGA_08	0.422	−0.324	0.465	−0.732*	0.458	0.719*	0.324	−0.317	SAR406
MGA_130	0.526	−0.444	0.570*	−0.754*	0.563	0.786*	0.444	−0.338	SAR406
MGA_131	0.782*	−0.746*	0.725*	−0.570*	0.669*	0.602*	0.746*	−0.697*	A714018
MGA_50	0.724*	−0.676*	0.725*	−0.725*	0.739*	0.786*	0.676*	−0.521	A714018
MGA_121	0.330	−0.317	0.349	−0.918**	0.680*	0.878**	0.317	−0.235	A714018

Abbreviations: MGA, Marine Group A; NESAP, Northeast subarctic Pacific Ocean; OTU, operational taxonomic unit.
*$P<0.05$; **$P<0.000079$, Bonferroni corrected.

However, when correlations were calculated for each station independently, statistically significant correlations were only identified at station P26 where MGA abundance was more strongly correlated with decreasing O_2 and increasing nitrate and phosphate concentrations than with temperature, Chl*a* or silicate (Table 2, Supplementary Figure S5). When calculated across the entire transect and each station independently, the relative abundance of MGA OTUs based on 16S rRNA gene clone library sequences was not significantly correlated with environmental parameters (data not shown). However, the relative abundance of four OTUs identified in pyrotags showed significant correlations across the entire transect with decreasing O_2 after a Bonferroni correction was applied ($P<0.000079$; Table 3). OTUs significantly correlated with decreasing O_2 were affiliated with two subgroups of MGA (Arctic95A-2 and A714018), and an additional 13 OTUs affiliated with HF770D10, ZA3648c, Arctic96B-7, Arctic95A-2, SAR406 and A714018 were weakly correlated ($P<0.05$; Table 3). In addition, out of all 78 MGA OTUs identified by binning pyrotags to full-length 16S rRNA gene sequences, 10 displayed significant correlations ($P<0.000079$) with increasing depth, salinity and nutrients (nitrate, phosphate, silicate) or decreasing Chl*a* (Supplementary Table S3).

Discussion

MGA abundance estimates in the NESAP were highly correlated between CARD-FISH and 16S rRNA gene clone library, but not between CARD-FISH and pyrotag sequences, whereas 16S rRNA gene clone library and pyrotag sequences were correlated based on Spearman's rank correlations. Moreover, CARD-FISH-based estimates were consistently <16S rRNA gene clone library or pyrotag sequence estimates for the same samples. For example, the average relative abundance of MGA sequences in O_2-deficient waters was 11.0% ± 3.9% based on 16S rRNA gene clone libraries, 9.9% ± 1.8% based on pyrotags and 5.6% ± 1.9% based on CARD-FISH (Table 1). This suggests an under or overestimation of MGA abundance by one, some or all of the methods used. The discrepancy between methods could be purely based on primer and probe differences and the underlying methods applied. One perspective would be that CARD-FISH with probe SAR406–97 underestimated MGA abundance. Lower detection efficiency by CARD-FISH has been attributed to limited probe access to target cells when using horseradish peroxidase-labeled probes (Schoenhuber *et al.*, 1997), even after careful permeabilization optimization (Woebken *et al.*, 2007). Also, the permeabilization step might cause leakage of ribosomes from target cells, which in turn could result in low-ribosome content cells dropping below the CARD-FISH detection limit (Hoshino *et al.*, 2008). Alternatively, MGA subgroups could harbor variable copy numbers of the 16S rRNA gene, inflating PCR-based metrics (Acinas *et al.*, 2004).

Rarefaction curves for MGA 16S rRNA gene clone library and pyrotag sequences recovered from the NESAP were consistent with known methodological limitations based on variable sample size and potential primer bias (Engelbrektson *et al.*, 2010; Schloss *et al.*, 2011; Gihring *et al.*, 2012). Clustering the combined data sets enabled pyrotag assignments to 78 out of 121 OTUs defined by 16S rRNA gene clone library sequences. The inability to assign pyrotags to all 121 OTUs may have resulted from

the conservative nature of our clustering method: we required full-length pyrotag sequences to match a cognate 16S rRNA gene clone library sequence with no mismatches. Alternatively, it is possible that time variable patterns in the abundance of MGA OTUs prevented assignment of all June 2009 pyrotags to OTUs identified in February 2009 16S rRNA gene clone libraries. Although ~50–75% of pyrotags identified as MGA in blast-based taxonomic queries were not assigned to OTUs defined by 16S rRNA gene clone library sequences, pyrotags affiliated with all 10 MGA subgroups were recovered. Indeed, comparison of 16S rRNA gene clone library and pyrotag sequence clusters revealed that the majority of MGA sequences (57%) and abundant MGA OTUs (containing >200 sequences) were identified using both methods (Figure 5). Unique pyrotag OTUs were generally composed of <50 sequences with a single abundant OTU containing 1490 pyrotags that could not be assigned to defined MGA subgroups. Sequences in this OTU were recovered from 500 m, 1000 m, 1300 m and 2000 m samples at all three stations indicating an environmental origin. The extent to which unique pyrotag OTUs captured components of the 'rare biosphere' (Sogin *et al.*, 2006) subject to time-variable changes in population structure remains to be determined. Despite this uncertainty, the recovery of a single abundant OTU unaffiliated with MGA subgroups defined by 16S rRNA gene clone library sequences suggests that the majority of abundant MGA subgroups in the NESAP have been identified.

Spearman's rank correlation coefficients provided statistical support for vertical partitioning of MGA subgroups in the NESAP water column. The relative abundance of MGA OTUs identified in pyrotags (affiliated with Arctic95A-2 and Arctic96B-7) displayed a negative correlation with O_2 concentration consistent with habitat selection within suboxic waters (1–20 $\mu mol\,kg^{-1}$) of the OMZ. The extent to which patterns of vertical partitioning among and between MGA OTUs represent ecological types (ecotypes; Koeppel *et al.*, 2008) or class divisions remains to be determined. Environmental gradients are common drivers of selection among microorganisms at different ecological scales. For example, Johnson *et al.* (2006) documented niche partitioning of *Prochlorococcus* ecotypes over ocean-basin scales across temperature (eMED4 vs eMIT9312) and nutrient (eNATL2A or eMIT9313) gradients. Similarly, SAR11 ecotypes display depth-specific distributions with subclade Ia members more prevalent in the euphotic zone and subclade II members more abundant in deeper (mesopelagic) waters (Field *et al.*, 1997). Such distribution patterns are associated with changes in genome composition that promote differential fitness including allelic variation (Urbach and Chisholm, 1998; Urbach *et al.*, 1998; Wilhelm *et al.*, 2007; Zhao and Qin, 2007) and metabolic island formation (Rocap *et al.*, 2003; Coleman *et al.*, 2006; Coleman and Chisholm, 2007; Kettler *et al.*, 2007;Wilhelm *et al.*, 2007).

Looking forward, genome-scale sequence data (that is, single-cell and metagenomic data) representative of defined MGA subgroups will be invaluable both to more accurately assess evolutionary relationships between MGA and thermophilic bacteria, such as *Calditrhrix*, as well as to attach metabolic repertoires to defined MGA subgroups (Shapiro *et al.*, 2012; Swan *et al.*, 2011). In turn, metabolic characterization of MGA subgroups will assist in determining whether observed 16S rRNA-based patterns of distribution across the oxycline are associated with variable forms of energy metabolism, consistent with redox-driven niche partitioning and ecotype differentiation. In addition, more extensive quantitative studies documenting the temporal dynamics of extant MGA subgroups across multiple provinces are needed to assess the stability of MGA population structure and function and better constrain the ecological and biogeochemical roles of MGA within OMZs.

Accession numbers

Bacterial 16S rRNA sequences reported in this study were deposited in GenBank with the accession numbers HQ242143–HQ242376 and HQ671746–HQ674628. Bacterial 16S rRNA sequences previously published in Walsh *et al.* (2009b) can be found under the accession numbers GQ351133–GQ351265. Pyrotag sequences reported in this study were deposited in GenBank with the accession number SRA051605.

Acknowledgements

We thank scientists and crew aboard CCGS *John P Tully* and the Canadian Department of Fisheries and Oceans for logistical support, in particular Kendra Mitchell, Olena Shevchuck, Karl Schiffmacher and Marie Robert; Bonnie Poulos and the Arizona Research Labs Cytometry Core Facility funded in part by the Cancer Center Support Grant (CCSG CA 023074) for assistance in sample analysis; Martha Schattenhofer and Bernhard Fuchs for providing positive controls for probe SAR406-97 and fruitful discussions; Niels Hanson for assistance with data visualization; Mike Gura, Stratis Gavaris and Evan Durno for assistance with statistical interpretation; and all members of the Tucson Marine Phage Lab and Hallam Lab for helpful comments along the way. We also thank the Joint Genome Institute, including Natasha Zvenigorodsky, Stephanie Malfatti, Phil Hugenholtz, Susan Yilmaz and Tijana Glavina del Rio for technical and project management assistance. This work was performed under the auspices of the US Department of Energy Joint Genome Institute under contract no. DE-AC02–05CH11231, the Natural Sciences and Engineering Research Council (NSERC) of Canada, Canada Foundation for Innovation and the Canadian Institute for Advanced Research through grants awarded to SJH as well as BIO5, Biosphere2, and NSF (OCE-0961947) grants awarded to MBS. JJW was supported by NSERC.

References

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* **186**: 2629–2635.

Amann R, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* **56**: 1919–1925.

Bano N, Hollibaugh JT. (2002). Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Appl Environ Microbiol* **68**: 505–518.

Barwell-Clarke J, Whitney F. (1996). IOS nutrient methods and analysis. *Can Tech Rep Hydrog Ocean Sci* **182**: 1–43.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

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.

Coleman ML, Chisholm SW. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* **15**: 398–407.

Cuttelod A, Herve C. (2010). ALMOFRONT 2 cruise in Alboran sea: Chlorophyll fluorescence calibration. *J Oceanogr Res Data* **3**: 6–11.

Daims H, Brühl A, Amann R, Schleifer K-H, Wagner M. (1999). The domain-specific probe EUB338 is insufficient for the detection of all *bacteria*: development and evaluation of a more comprehensive probe set. *Syst Appl Microbiol* **22**: 434–444.

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.* (2006a). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

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

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H *et al.* (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* **4**: 642–647.

Field KG, Gordon D, Wright T, Rappe M, Urback E, Vergin K *et al.* (1997). Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl Environ Microbiol* **63**: 63–70.

Freeland H. (1997). A short history of Ocean Station Papa and Line P. *Prog Oceanogr* **75**: 120–125.

Fuchs BM, Woebken D, Zubkov MV, Burkill P, Amann R. (2005). Molecular identification of picoplankton populations in contrasting waters of the Arabian Sea. *Aquat Microbiol Ecol* **39**: 145–157.

Fuhrman JA, McCallum K, Davis AA. (1993). Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Appl Environ Microbiol* **59**: 1294–1302.

Fuhrman JA, Davis AA. (1997). Widespread archaea and novel bacteria from the deep sea as shown by 16S rRNA gene sequences. *Mar Ecol Prog Ser* **150**: 275–285.

Gihring TM, Green SJ, Schadt CW. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14**: 285–290.

Gordon D, Giovannoni S. (1996). Detection of stratified microbial populations related to *Chlorobium* and *Fibrobacter* species in the Atlantic and Pacific Oceans. *Appl Environ Microbiol* **62**: 1171–1177.

Guindon S, Lethiec F, Duroux P, Gascuel O. (2005). PHYML online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33**: W557–W559.

Helly JJ, Levin LA. (2004). Global distribution of naturally occurring marine hypoxia on continental margins. *Oceanogr Res Pap* **51**: 1159–1168.

Holm-Hansen O, Lorenzen CJ, Holmes RW, Strickland JDH. (1965). Fluorometric determination of chlorophyll. *J Cons Perm Int Explor Mer* **30**: 3–15.

Hoshino T, Yilmaz LS, Noguera DR, Daims H, Wagner M. (2008). Quantification of target molecules needed to detect microorganisms by fluorescence *in situ* hybridization (fish) and catalyzed reporter deposition-FISH. *Appl Environ Microbiol* **74**: 5068–5077.

Huber T, Faulkner G, Hugenholtz P. (2004). Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**: 2317–2319.

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

Kettler GC, 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.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM *et al.* (2008). Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *Proc Natl Acad Sci* **105**: 2504–2509.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118–123.

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

Madrid VM, Taylor GT, Scranton MI, Chistoserdov AY. (2001). Phylogenentic diversity of bacterial and archeal communities in the anoxic zone of the Cariaco Basin. *Appl Environ Microbiol* **67**: 1663–1674.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A *et al.* (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.

Miroshnichenko ML, Kostrikina NA, Chernyh NA, Pimenov NV, Tourova TP, Antipov AN *et al.* (2003). *Caldithrix abyssi* gen. nov., sp nov., a nitrate-reducing, thermophilic, anaerobic bacterium isolated from a Mid-Atlantic Ridge hydrothermal vent, represents a novel bacterial lineage. *Int J Syst Evol Microbiol* **53**: 323–329.

Pace NR. (2009). Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev* **73**: 565–576.

Paulmier A, Ruiz-Pino D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. *Prog Oceanogr* **80**: 113–128.
Pena MA, Bograd SJ. (2007). Time series of the northeast Pacific. *Prog Oceanogr* **75**: 115–119.
Pena MA, Varela DE. (2007). Seasonal and interannual variability in phytoplankton and nutrient dynamics along Line P in the NE subarctic Pacific. *Prog Oceanogr* **75**: 200–222.
Pernthaler A, Pernthaler J, Amann R. (2004). Sensitive multi-color fluorescence *in situ* hybridization for the identification of environmental microorganisms. In: Kowalchuk GA, De Bruijn FJ, Head IM, Akkermans AD, van Elsas JD (eds). *Molecular Microbial Ecol Man*, 2nd edn. Kluwer Academic Publishers: Dordrecht, pp 711–726.
Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.
Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.
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.
Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pernthaler J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ Microbiol* **11**: 2078–2093.
Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
Schloss PD, Gevers D, Westcott SL. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**: e27310.
Schoenhuber W, Fuchs BM, Juretschenko S, Amann R. (1997). Improved sensitivity of whole-cell hybridization by the combination of horseradish peroxidase-labeled oligonucleotides and tyramide signal amplification. *Appl Environ Microbiol* **63**: 3268–3273.
Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G *et al.* (2012). Population genomics of early events in the ecological differentiation of Bacteria. *Science* **336**: 48–51.
Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci* **103**: 12115–12120.
Stevens H, Ulloa O. (2008). Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environ Microbiol* **10**: 1244–1259.
Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D *et al.* (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**: 1296–1300.
Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I *et al.* (2010). 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytol* **188**: 291–301.
Urbach E, Chisholm SW. (1998). Genetic diversity in *Prochlorococcus* populations flow cytometrically sorted from the Sargasso Sea and Gulf Stream. *Limnol Oceanogr* **43**: 1615–1630.
Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW. (1998). Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (*Cyanobacteria*). *J Mol Evol* **46**: 188–201.
Wallner G, Amann R, Beisker W. (1993). Optimizing fluorescent *in situ*-hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry* **14**: 136–143.
Walsh DA, Zaikova E, Hallam SJ. (2009a). Large volume (20L+) filtration of coastal seawater samples. *J Vis Exp* **28**: 1161.
Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG *et al.* (2009b). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**: 578–582.
Whitney FA, Freeland HJ, Robert M. (2007). Persistently declining oxygen levels in the interior waters of the eastern subarctic Pacific. *Prog Oceanogr* **75**: 179–199.
Whitney FA, Wong CS, Boyd PW. (1998). Interannual variability in nitrate supply to surface waters of the northeast pacific ocean. *Mar Ecol Prog Ser* **170**: 15–23.
Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**: 27.
Woebken D, Fuchs BM, Kuypers MMM, Amann R. (2007). Potential interactions of particle-associated anammox bacteria with bacterial and archaeal partners in the Namibian upwelling system. *Appl Environ Microbiol* **73**: 4648–4657.
Wright JJ, Lee S, Zaikova E, Walsh DA, Hallam SJ. (2009). DNA extraction from 0.22 microM Sterivex filters and cesium chloride density gradient centrifugation. *J Vis Exp* **31**: pii 1352.
Wright JJ, Konwar KM, Hallam SJ. (2012). Microbial ecolgy of expanding oxygen minimum zones. *Nat Rev Microbiol* **10**: 381–394.
Zaikova E, Hawley A, Walsh DA, Hallam SJ. (2009). Seawater sampling and collection. *J Vis Exp* **28**: e1159.
Zaikova E, Walsh DA, Stilwell CP, Mohn WW, Tortell PD, Hallam SJ. (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ Microbiol* **12**: 172–191.
Zhao FQ, Qin S. (2007). Comparative molecular population genetics of phycoerythrin locus in Prochlorococcus. *Genetica* **129**: 291–299.

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