RESEARCH ARTICLE

Ecological structuring of bacterial and archaeal taxa in surface ocean waters

Pelin Yilmaz[1,2], Morten H. Iversen[3], Wolfgang Hankeln[1,2], Renzo Kottmann[1], Christian Quast[1] & Frank O. Glöckner[1,2]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany; [2]Jacobs University Bremen gGmbH, Bremen, Germany; and [3]Faculty of Geosciences and MARUM, University of Bremen, Bremen, Germany

Correspondence: Frank O. Glöckner, Max Planck Institute for Marine Microbiology, Celsiusstr. 1, 28359 Bremen, Germany. Tel.: +49 421 2028 970; fax: +49 421 2028 580; e-mail: fog@mpi-bremen.de

Received 22 June 2011; revised 2 December 2011; accepted 4 March 2012.

DOI: 10.1111/j.1574-6941.2012.01357.x

Editor: Patricia Sobecky

Keywords
microbial community structure; bacterioplankton; Global Ocean Sampling expedition; ontology; marine habitats; ribosomal RNA.

Abstract

The Global Ocean Sampling (GOS) expedition is currently the largest and geographically most comprehensive metagenomic dataset, including samples from the Atlantic, Pacific, and Indian Oceans. This study makes use of the wide range of environmental conditions and habitats encompassed within the GOS sites in order to investigate the ecological structuring of bacterial and archaeal taxon ranks. Community structures based on taxonomically classified 16S ribosomal RNA (rRNA) gene fragments at phylum, class, order, family, and genus rank levels were examined using multivariate statistical analysis, and the results were inspected in the context of oceanographic environmental variables and structured habitat classifications. At all taxon rank levels, community structures of neritic, oceanic, estuarine biomes, as well as other exotic biomes (salt marsh, lake, mangrove), were readily distinguishable from each other. A strong structuring of the communities with chlorophyll *a* concentration and a weaker yet significant structuring with temperature and salinity were observed. Furthermore, there were significant correlations between community structures and habitat classification. These results were used for further investigation of one-to-one relationships between taxa and environment and provided indications for ecological preferences shaped by primary production for both cultured and uncultured bacterial and archaeal clades.

Introduction

Ecological structuring of *Bacteria* and *Archaea* from a range of habitats, at genera or even species level, is nowadays routinely investigated (Lauber *et al.*, 2009; Andersson *et al.*, 2010; Kirchman *et al.*, 2010). For the human nature, it is tempting to characterize and categorize ‘objects’, and assign them to ‘containers’ – big or small – which reflect particular characteristics of all these objects (Philippot *et al.*, 2010). Still, there is controversial debate about bacterial and archaeal ecologically coherent containers, mainly because of the vast genetic and physiological diversity contained at high-level ranks. However, there is also striking evidence that correlations between taxonomy and functions exist. For example, several studies were able to associate phyla or classes with r- or K-type life strategies; in marine systems, members of SAR11 and *Bacteroidetes* were identified as K-strategists, and in soil systems, *Betaproteobacteria* were found to be r-strategists (Alonso-Sáez *et al.*, 2006; Fierer *et al.*, 2007). Other studies demonstrated specific taxa–habitat associations, via either cross- or within-habitat comparisons (Glöckner *et al.*, 1999; von Mering *et al.*, 2007; Nemergut *et al.*, 2010). Finally, investigations of responses of specific taxa to changing environmental conditions showed supportive results (Fuhrman *et al.*, 2006; Pommier *et al.*, 2007; Philippot *et al.*, 2009).

The Global Ocean Sampling (GOS) (Rusch *et al.*, 2007; Yooseph *et al.*, 2010) provides a range of marine and aquatic habitats, enabling both inter- and intrahabitat comparisons. The sequences are associated with a relatively rich set of associated data (contextual or metadata), such as geographic coordinates and environmental variables, thus making the GOS expedition suitable for a meta-analysis. This study explores the possible ecological cohesions in high-level taxa of surface ocean *Bacteria* and *Archaea*, using taxonomically classified 16S ribosomal RNA gene fragments (rRNA) from the GOS metagenomes. We evaluated the results in the framework of comparing high-level taxa ranks to low-level taxa, annotating sampling sites with ontological habitat classifications (Environment Ontology, http://www.environmentontology.org) at three different granularity levels (biome, feature, and material), and correlating community structures, as well as relative abundances of selected marine taxa to directly measured and inferred environmental factors.

Materials and methods

Retrieval and alignment of SSU rRNA fragments

Unassembled metagenomic reads for 80 GOS sample datasets were downloaded as a FASTA file from the CAMERA website (Seshadri *et al.*, 2007) on September 2009. A total of 10 085 737 reads, with an average read length of 822 bp, were processed with a custom-tailored configuration of the SILVA pipeline (Pruesse *et al.*, 2007) in order to retrieve SSU rRNA gene fragments.

Firstly, a quality inspection was conducted. Reads composed of more than 2% of ambiguous bases or more than 2% of homopolymeric stretches longer than four bases were rejected. Additionally, reads having more than 5% identity to vector sequences based on BLASTN hits were excluded (Altschul *et al.*, 1990). The database used for vector contamination checking was a combined vector sequence database based on the EMVEC (ftp://ftp.ebi.ac.uk/pub/databases/emvec/) and UNIVEC (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Secondly, the SILVA INcremental Aligner (SINA) was used to spot and align actual SSU rRNA gene fragment regions (Pruesse *et al.*, 2011). Finally, the sequences were imported into ARB (Ludwig *et al.*, 2004) for further analysis.

Taxonomic classification of fragments

Aligned SSU rRNA gene fragments were added to the guide tree included in the SSU dataset of the SILVA Reference (Ref) release 104 using the ARB Parsimony tool. Fragments having 100–500 bases within the rRNA gene boundaries were added to the guide tree using individual *Bacteria*, *Archaea*, and *Eukarya* filters, excluding highly variable positions between 1 and 7 leaving 1391 of 1444 positions. For fragments with more than 500 aligned bases, sequences were added with the same positional variability filters but excluding highly variable positions between 1 and 9 leaving 1224 valid positions. Taxonomic assignments were based on membership of the fragments to the existing clades of the SILVA taxonomy. Manual refinement of the taxonomic groups was performed after the addition of all GOS rRNA gene sequences. Taxonomic path assignments were stored in the 'tax_slv' field of ARB files using the taxonomy(n) function of ARB Command Interpreter.

Environmental parameters and habitat assignments

Temperature, salinity, and chlorophyll *a* concentration values were taken from *in situ* measurements, when available (Rusch *et al.*, 2007). Dissolved oxygen, nitrate, phosphate, and silicate concentrations were interpolated from World Ocean Atlas 2005 (WOA05) and World Ocean Database 2005 (WOD05) data at the geographic locations using the GIS tools of the megx.net portal (Kottmann *et al.*, 2010) (Supporting Information, Table S1). The usage of interpolated nutrient values in microbial ecology studies is a relatively new practice, but this has been applied in several studies to date (Gianoulis *et al.*, 2009; Luo *et al.*, 2009; Martiny *et al.*, 2009; Kostadinov *et al.*, 2011; Temperton *et al.*, 2011). In order to provide a solid example of the usefulness of WOA05 data, we extracted bottle data from HOT-ALOHA (2) station for these parameters, for the time period between 2000–2010, and depth interval of 0–5 m, and calculated monthly averages for this time period (Table S2 and Fig. S1). These averages were then compared to monthly-interpolated values from the WOA05 for the same coordinates. Despite discrepancies observed at certain months for nitrate, phosphate, and silicate, the values from both data sources could be considered a good match, and in the absence of *in situ* data, WOA05 interpolation is a useful alternative.

World ocean net primary productivity map was generated by integrating annual average net primary productivity estimations using the standard vertically generalized production model (http://www.science.oregonstate.edu/ocean.productivity/index.php) (Behrenfeld & Falkowski, 1997) from 2004 into megx.net and using custom map generation configuration for visualization.

Habitat assignments for GOS sites were manually curated using an edit version of Environment Ontology (EnvO) terms (http://obo.cvs.sourceforge.net/viewvc/obo/obo/ontology/environmental/envo-edit.obo), at three different levels as biome, feature, and material (Table S3).

Statistical analysis

Absolute abundances of taxa were standardized using the mean abundance of a set of 'single-copy domains' (SCDs), namely b5, ef_ts, pnpase, rbfa, rrf, ribosomal_l12, ribosomal_l15, ribosomal_l16, ribosomal_l19, ribosomal_l20, ribosomal_l21p, ribosomal_l27, ribosomal_l29, ribosomal_l9_n, ribosomal_s16, ribosomal_s20p, ribosomal_s3_c, ribosomal_s3_n, srp_spb, smpb, upf0054, trna_m1g_mt. These SCDs were found to occur only once in a set of 43 completely sequenced genomes of both marine and nonmarine isolates (Table S4). GOS metagenomes were queried by hidden Markov models belonging to these SCDs present in the PFAM 23.0 database (Finn *et al.*, 2008) using a single TimeLogic DeCypher card (Active Motif, Inc., Carlsbad, CA), and hits with E-values below 10^{-10} were used for the standardization. The standardized absolute abundances were then converted to relative abundances by dividing them by the total count of all 16S rRNA gene fragments with more than 100 bases at each GOS site (Table S5).

Following the standardization, relative abundances were converted into a sites*species matrix, complemented by a sites*parameters matrix, and imported into the R statistical computing environment (R Development Core Team, 2010) for further statistical analysis and visualization purposes.

The R package vegan v1.17-3 (Oksanen *et al.*, 2011) was used for all numerical ecology analyses. Bray–Curtis dissimilarities were calculated between GOS sites based on Wisconsin and square-root-standardized sites*species matrices of different taxonomic ranks levels (phylum, class, order, family, genus), and used in the metaMDS nonmetric dimensional scaling (NMDS) procedure. Taxa–environment relationships were studied using least squares linear vector fitting, after the variables were subjected to z-score standardization. For categorical environmental variables, or factors (habitat assignments), centroids (average scores) with standard deviations were calculated. Significance of the fitted vectors or factors was determined by permutations ($n = 9999$), and a Pr (> r) value < 0.01 was judged to be significant. Additionally, a generalized additive model surface fit was visualized as smooth, nonparametric isoclines with significance tested by permutation tests ($n = 9999$) and ANOVA. The coefficient of determination (R2) was used as a goodness-of-fit measure for fitted vectors, factors, and nonparametric surfaces.

Data access

16S rRNA sequences retrieved from the GOS metagenomes that were analyzed in this study are publicly available from www.arb-silva.de/download/archive/GOS_diversity/ in ARB format, as well as unaligned and aligned FASTA files. The.arb file contains the SILVA_Ref_104 guide tree with GOS 16S rRNA fragments added by parsimony, and the.fasta contains only GOS 16S rRNA fragments with their taxonomic assignments included in the FASTA header.

Results and discussion

Overview of the taxonomic makeup

The quantity and length distribution of the retrieved 16S rRNA gene fragments were described elsewhere (Yilmaz *et al.*, 2011), while the number of taxa at each rank level, overall taxonomic composition, and other details regarding the taxonomic classification can be found in the supplementary material (Table S6, S7, and S8).

The overall assessment of the taxonomic makeup is in agreement with previous studies on the GOS metagenome (Rusch *et al.*, 2007; Biers *et al.*, 2009; Yooseph *et al.*, 2010), as well as with expectations of ocean surface microbial communities (Fuhrman & Hagström, 2008). Compared to previous assessments of the GOS metagenome taxonomic makeup, we have observed more sequences belonging to the Candidate divisions. For example, only Candidate division OD1 is acknowledged in the study by Biers and colleagues (Biers *et al.*, 2009), whereas we report the occurrence of TM7, WS3, OP3, SR1, and OP10. Clone sequences belonging to these divisions are isolated from a wide variety of sources, including sludge, soil, human or other host tissues, deep-sea sediments, lakes, or biofilms (Hugenholtz *et al.*, 1998; Pace, 2009). In the GOS metagenome, their distribution was limited to, except for Candidate division OD1, coastal, estuarine, brackish, hypersaline waters, as well as freshwater environments. The absence of these divisions from surface open ocean waters is congruent with previous observations, whereas the presence in estuarine and coastal waters could be indicative of anthropogenic inputs considering their prevalence in wastewater/sludge type environments, while the fresh, brackish, and hypersaline water prevalence is in line with potentially differing metabolic capabilities in comparison with surface ocean communities. Candidate division OD1 was the most widespread candidate division within the GOS metagenome, and in addition to the previously listed locations, this taxon was also observed in ocean waters from sites GS000a, GS114, and GS117 (Fig. S2). Although the OD1 is environmentally widespread, our survey of previous isolation sources did not encounter any other surface ocean clones.

Community structures at different taxonomic rank levels

Spatial and temporal patterns, along with ecological coherence of higher bacterial and archaeal taxonomic ranks, have been discussed previously (Philippot *et al.*, 2009, 2010). Although the GOS metagenome is only composed of surface water samples, the 'surfaces' sampled have an interesting variety of contrasting habitats, such as estuary vs. open ocean, or hypersaline vs. freshwater. We used this diversity of habitats in order to reveal how well these habitat differences will be reflected in community structures composed of bacterial and archaeal taxa at different rank levels.

The sites*species matrix for ordination analysis consisted of standardized relative abundances at five different rank levels. The phylum level consisted of 33 distinct taxa, with 12 313 sequences classified into these taxa. The class level had 55 distinct taxa, with 12 222 sequences; order level 107 distinct taxa and 12 049 sequences; family level 201 distinct taxa and 10 616 sequences; and finally genus level 363 distinct taxa and 4270 sequences.

At any taxonomic rank level, the NMDS analysis showed that certain sites have remarkably different community structures (Fig. 1). Specifically, a recurring trend was a halo of coastal (GS013), estuary (GS011–GS012), hypersaline (GS033), freshwater (GS020), mangrove (GS032), fringing reef (GS025), and some open ocean (GS00a–c) sites, surrounding a cluster of mainly open ocean sites. In addition to the aforementioned sites, another set of coastal/estuary (GS002–GS010), warm seep (GS030), coral reef (GS048), and coastal upwelling (GS031) sites were to some extent distinguishable from the open ocean cluster. The relative distances between sites at different ranks were not always the same. For example, GS011–GS012 couple was placed at varying distances from each other, but nevertheless they retained their general distinctness from the rest. Another one is GS013 and GS025 couple, clearly different from the rest of the sites, but conspicuously placed too near each other at phyla rank level. The taxa composition and relative abundances change with each rank level; therefore, such conformation changes are expected. The NMDS ordinations have high stress values, although still being within acceptable limits (Kruskal, 1964). Other multivariate analysis methods may also be suitable; however, NMDS has the advantage of being a nonparametric method, therefore not assuming that species have linear responses to environmental gradients.

Because a systematic classification of the habitat types can extend these observations further, all NMDS plots were annotated with Environment Ontology (EnvO) biome, feature, and material terms (Fig. 2). The three different levels of EnvO terms provide an increasing order of granularity to habitat description of the sampling sites; the first-level biome (e.g. large lake or estuarine) broadly establishes the system that defines the scope of potential ecological inputs that a biological entity may be subjected to, whereas an environmental feature (e.g. atoll, bay) describes a range of biotic and abiotic entities and phenomena that are more local to that entity than its biome, and finally, material (e.g. coastal water, ocean water) is understood as the substance immediately surrounding that entity and acting as the primary transmitter of ecological forces to and from it.

All EnvO term levels produced significant correlations with the ordinations; however, biome and material, overall, produced 1.5–2 times higher correlations, compared to feature terms. Although contrasting biome or material types, such as lake vs. oceanic, or hypersaline vs. estuarine water, were distinguishable on all rank levels, meaningful associations of identical terms started to appear at the class rank and improved at lower rank levels. At phylum and class level, for example, oceanic epipelagic zone and neritic epipelagic zone biomes, or coastal water and (open) ocean water sites were intermixed, while order level on these two biomes was more separated. These two clusters were not clearly separated, and some neritic/coastal sites were overlapping with the oceanic cluster. This was observed because most of these sites were sampled around islands and were heavily mixed with open ocean waters. Therefore, although the ontologically correct annotation would be 'neritic epipelagic zone biome' or 'coastal water', the community composition resembles the open ocean.

The two estuarine biomes were placed together on all rank-level ordinations; however, a third one occurred with neritic biomes and separated from the former two. This deviation can be explained by the locality; the former two sites are located at Delaware Bay and Chesapeake Bay, respectively, whereas the latter site is listed as Bay of Fundy, which are drastically different estuaries owing to higher anthropogenic influences at Delaware and Chesapeake Bays (Lotze *et al.*, 2006). This suggests that sites with the same habitat type may show differences in high- and low-level taxon ranks owing to external influences and/or mixing of water masses.

A number of other biomes were sampled during the GOS expedition, namely marine coral reef, marine reef, warm seep, and marginal sea. The ordinations did not reveal these biomes as being different from oceanic and neritic sites, although it is known that specific groups of bacteria are known to be associated with corals (Rohwer *et al.*, 2002; Pantos *et al.*, 2003; Bourne & Munn, 2005), and a low similarity between Pacific Ocean and Caribbean Sea samples has been observed previously (Lee & Fuhrman, 1991). The possible explanation for this observation can be that the habitat annotations are misleading and that the prominent feature of these sites is being neritic or oceanic biomes, rather than being reef or marginal sea biomes. Another clue supporting this argument is recognized

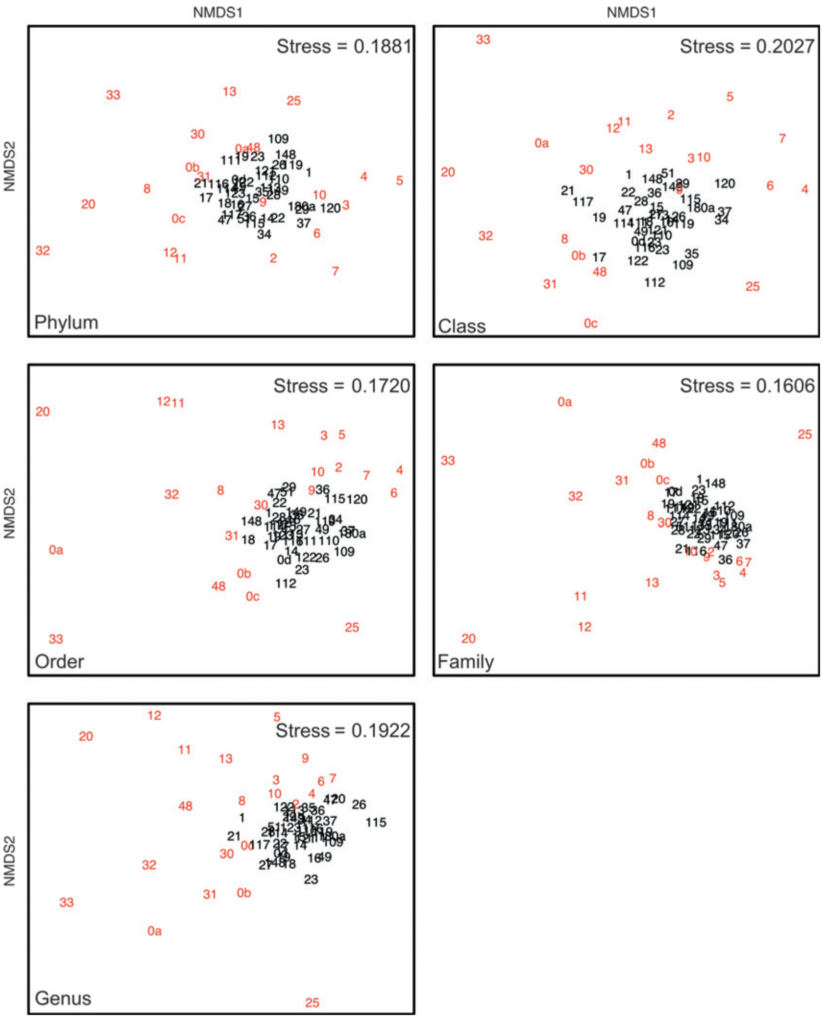

Fig. 1. Panel figure showing NMDS analysis for each taxonomic rank level. NMDS is an iterative search for ranking and placement of *n* entities (samples) in *k* dimensions (ordination axes) that minimizes the stress of the *k*-dimensional configuration. The 'stress' value is a measure of departure from monotonicity in the relationship between the dissimilarity (distance) in the original *p*-dimensional space and that in the reduced *k*-dimensional ordination space (Clarke, 1993; Ramette, 2007). NMDS is therefore used to find a configuration in a given number of dimensions, which preserves rank-order dissimilarities in species composition as closely as possible, such that distance along a NMDS axis corresponds to relative difference in community composition. The axes (NMDS1 and NMDS2) are arbitrary and just represent a framework for sampling site points; however, they are scaled so that one unit means doubling of community dissimilarity. The community dissimilarities were calculated based on taxa-standardized relative abundances. For visibility, the 'GS' prefix was omitted from sampling site names. Stress values are indicated at the top-right corner of each figure, whereas the ranks are indicated at the bottom-left corner. Sampling sites mentioned in Results and discussion are highlighted in red.

with EnvO feature annotations; the sample feature of the majority of these biomes is photic zone, and with this annotation, they are clustered with sites sharing this feature.

In summary, with the application of ontological annotations to sampling sites, a context to community structure differences was gained, whereby an understanding of ecological structuring of the high-level taxa can be observed. Our *in silico* observations, along with previous *in situ* and *in silico* evidence (Fierer *et al.*, 2007; Philippot *et al.*, 2009; Zinger *et al.*, 2011), support that higher taxonomic rank levels such as phylum or class provide enough information to distinguish between highly contrasting habitat types. Hence, phyla or classes can be used as indicator taxa to identify the specific habitats. However, at the interface of two habitats, like coastal vs. (open) ocean water, it is necessary to have more resolution for discriminatory power. A basic example is the case of *Betaproteobacteria*; at phylum level, this low brackish-preferring class will be accounted as *Proteobacteria*, hence

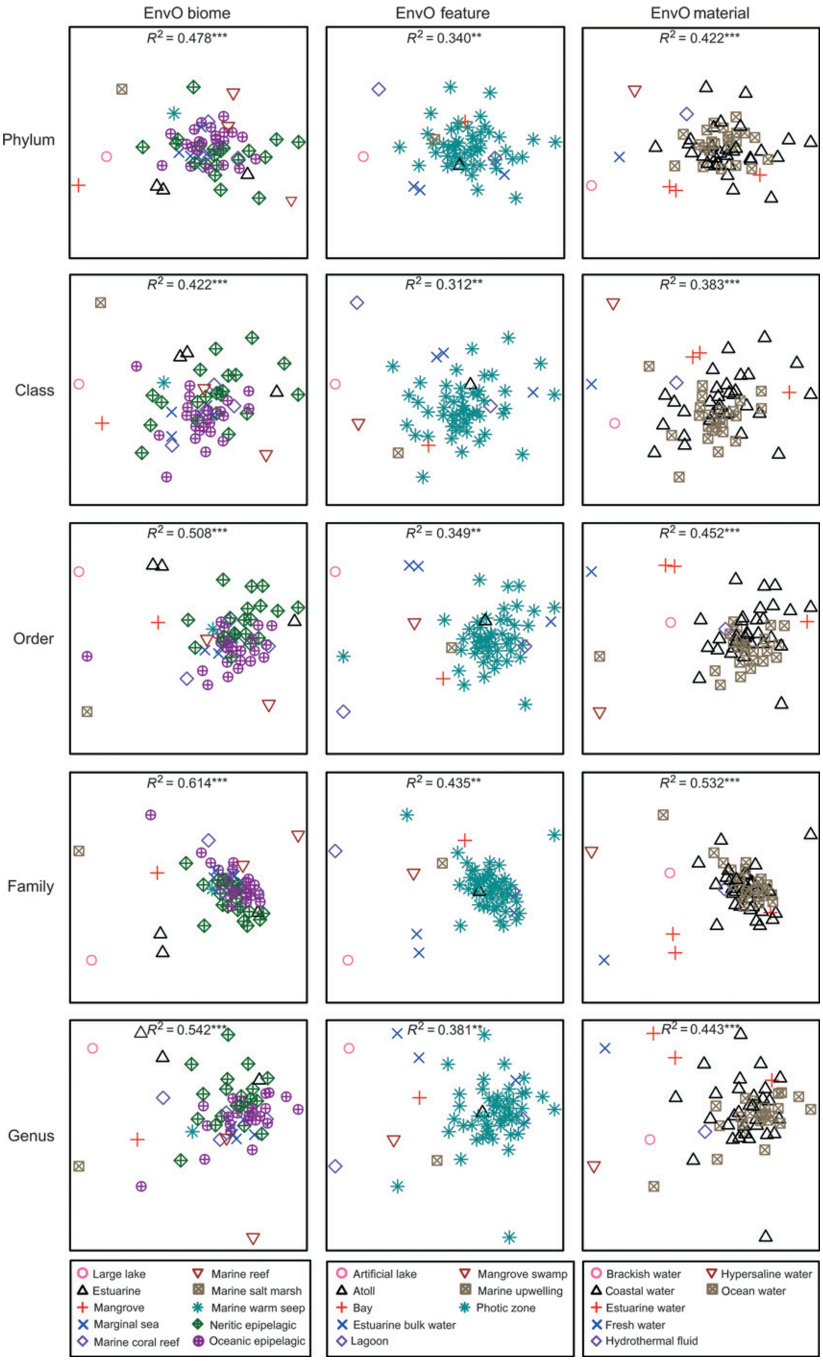

Fig. 2. Panel figure showing NMDS analysis for each taxonomic rank level. The configuration of sites is the same as Fig. 1; however, now sites are annotated with EnvO terms at three different levels, biome, feature, and material. Rows indicate different rank levels, whereas columns indicate different sets of terms. Legends at the bottom of each column show the color and shape code of EnvO terms. Goodness-of-fit of term levels to the ordination are indicated by the R2 values, and the significances by asterisk symbols ($0 < P$ *** $< 0.001 < P$ ** $< 0.01 < P$ * $< 0.05 < P < 0.1 < P < 1$).

leading to a poor ordination. Nevertheless, these ordinations do not provide clear-cut habitat clusters, but a certain amount of fuzziness is observed even at genus level.

To test whether bacterial and archaeal taxa distribution is related to environmental conditions, we fitted vectors and nonparametrically smoothed surfaces of seven environmental variables (Virtanen *et al.*, 2006), which were obtained both *in situ* and by interpolation. The combined interpretation of variable vectors and fitted surfaces is to be made as follows (see Fig. 3); the vector arrow points to the direction of most rapid change in the environmental variable, or the direction of the gradient, and the length of the arrow is proportional to the correlation between ordination and variable. A planar fitted surface indicates that the response of the community to the variable is linear, and the surface R2 will be equal to or close to the R2 of the vector. If the response is nonlinear, R2 for the surface will be higher than for the vector. For example, if the R2 values for temperature vector and fitted surface are equal, then this would imply that temperature has a direct effect on the bacterioplankton community, that is, by causing higher metabolic rates or death/dormancy in members of the community, hence changing the community structure. If the effect is nonlinear, then the effect of the environmental variable on the community structure will be indirect, that is, a high nutrient situation providing excess concentration of dissolved organic matter needed for bacterial growth via phytoplankton exudates (Larsson & Hagström, 1979).

Of the seven variables, only three, namely temperature, salinity, and chlorophyll *a* concentration, produced significant correlations with community structures (Fig. 3). It is surprising that the nutrients did not significantly correlate with the microbial community structure, as they have high contribution to the production of inorganic nutrients via remineralization in the surface waters (Azam *et al.*, 1983). Furthermore, it was suggested that silicate regeneration in the oceans is controlled by bacterial dissolution of diatom frustules (Bidle & Azam, 1999), implying there could be a link between silicate concentration and bacterioplankton community composition. Nevertheless, individual correlations of certain taxa with these nutrients cannot be dismissed, although they do not seem to affect the 'big picture'.

Temperature produced significant correlation at all rank levels except at phylum and class levels. The strength of the correlation was high at order level (0.394), but a small decrease at family level was observed (0.213), which was followed by an increase at genus level (0.313). In any case, the response of the community structure to temperature was nonlinear, as indicated by higher surface correlations. This finding both supports and contrasts previous studies; Pommier *et al.* (2007) and Fuhrman *et al.* (2008) showed that bacterial richness is linearly positively correlated with temperature and suggested a direct effect of temperature on bacterioplankton diversity through enzyme kinetics (Pommier *et al.*, 2007; Fuhrman *et al.*, 2008). Our results support that temperature is an important determinant of community composition, but because of nonlinear effects observed, secondary variables acting together with temperature are worth considering. For example, studies demonstrate that metabolic activity at low temperatures requires higher concentrations of specific substrates. Further, the temperature and respiration relationship, expressed as Q10, suggests an exponential relationship (Wiebe *et al.*, 1992; Pomeroy & Wiebe, 2001).

Salinity was a significant variable at phylum, order, and family levels, and correlation was highest at order level (0.417). As with temperature, the effect of salinity was also nonlinear. Again, a number of other studies have determined the importance of salinity on bacterial community composition, both on different aquatic environments (Nold & Zwart, 1998) and on global scale (Lozupone & Knight, 2007). The nonlinear effect observed concurs with the wide range of physiological effects that salinity can have (e.g. membrane potential, transport systems). Additionally, the indirect relationship could also be due to a secondary influence originating from mixing of water masses, which would both affect the salinity and community composition.

Chlorophyll *a* concentration correlations were significant at all levels, except at phylum level, and produced the strongest correlation of all the three variables. Additionally, a linear effect was observed at class level, although this effect changed to nonlinear at lower rank levels. As chlorophyll *a* concentration is an indicator of phytoplankton biomass, and as heterotrophic bacterioplankton depends on their products and remains, this is not an unexpected outcome. In fact, chlorophyll *a* concentration was found to be a determinant of seasonal and annual community composition dynamics (Fuhrman *et al.*, 2006; Gilbert *et al.*, 2011).

Temperature, salinity, and chlorophyll *a* concentrations are environmental variables with known effects on structuring the marine bacterial and archaeal communities. This study confirms those previous local observations on a global scale and underlines the environmental effects on ecological structuring of high taxon levels.

Controls on geographic distribution of marine clades

These correlations can provide novel indications about the relationships of variable gradients with taxa, which are useful, especially in the case of taxa with few or no cultured members. For example, clade BD1-5 and TM6 (phylum level) can be ascribed as brackish-preferring clades, whereas RF3 appears at extreme salinity levels (Fig. S2). However, clade BD1-5 was found both at GS033 (marine salt marsh biome) and at GS032 (low salinity with 29.47 PSU) as this clade occurs at both

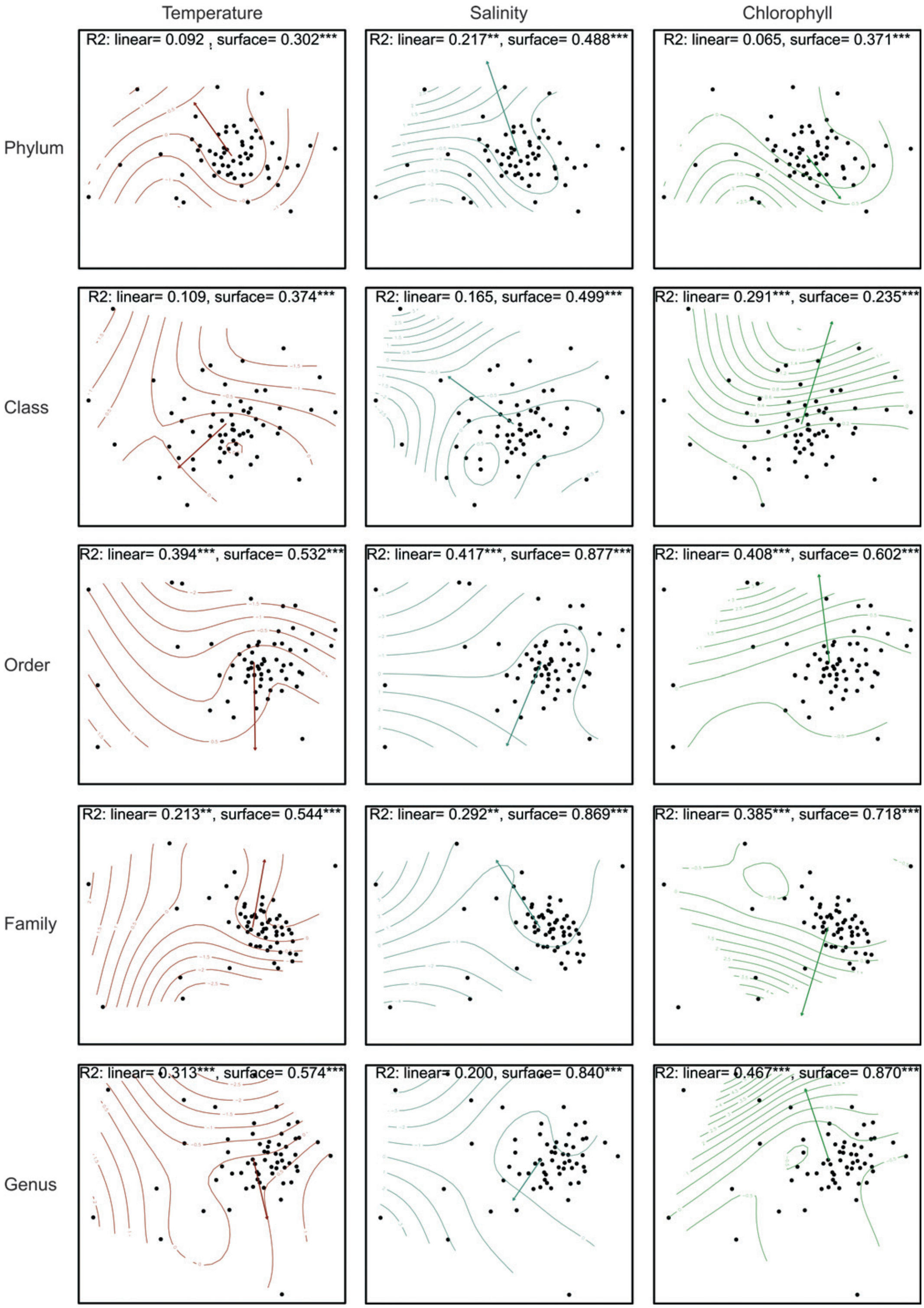

Fig. 3. Panel figure showing NMDS analysis for each taxonomic rank level with fitted environmental variable vectors and nonparametric surfaces. Rows indicate different rank levels, whereas columns indicate different variables. All variable values were z-score standardized prior to vector and surface fitting; hence, the isocline values reflect the z-scores. Goodness-of-fit of vectors and surfaces are again indicated by the R2 values, R2 linear, and R2 surface, respectively, while significances by asterisks (0 < *P* *** < 0.001 < *P* ** < 0.01 < *P* * < 0.05 < *P* < 0.1 < *P* < 1).

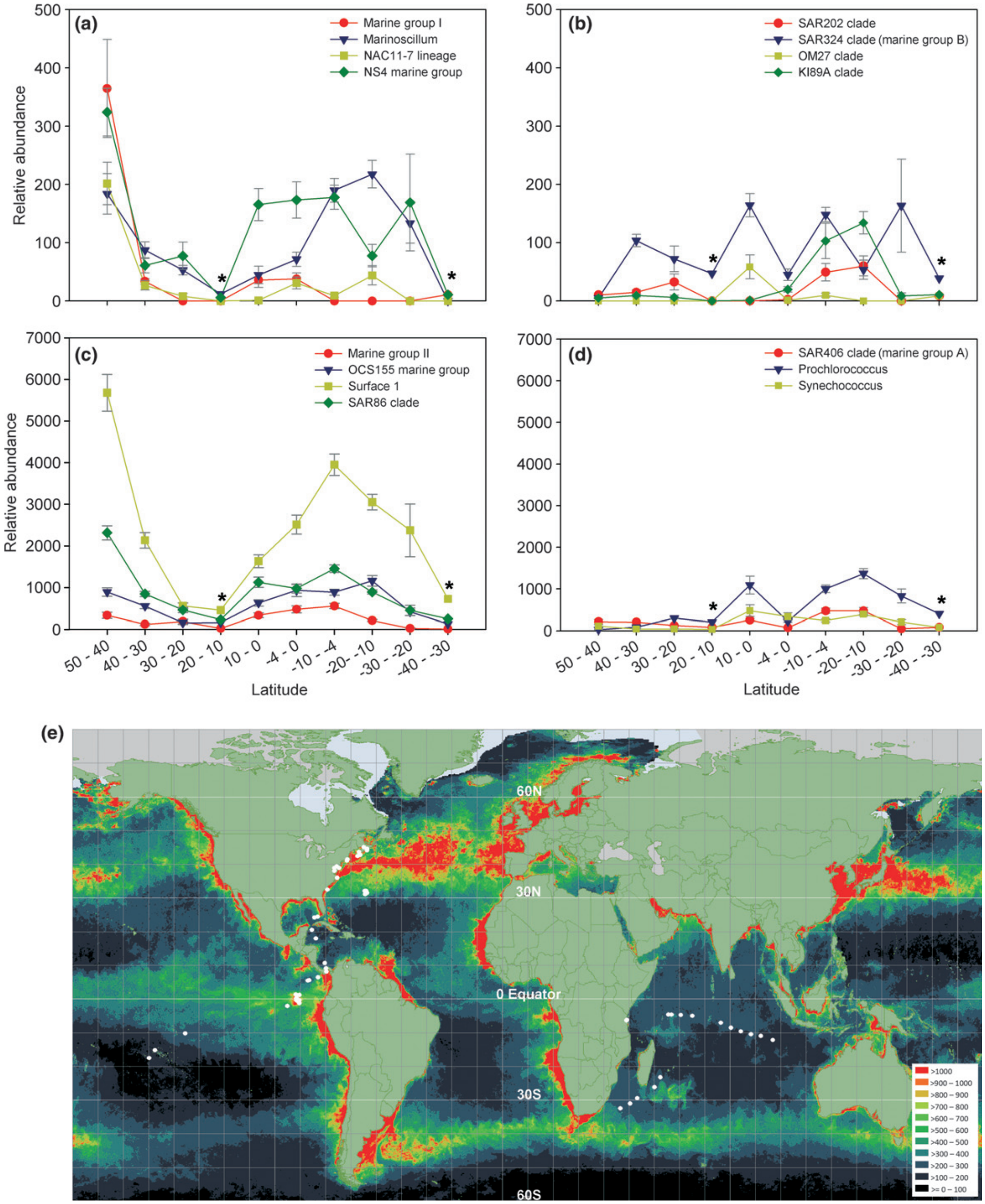

Fig. 4. Latitudinal distribution patterns for selected marine taxa. The relative abundances are calculated as described in the materials and methods section, and then by multiplying the resulting value with 10^5. Relative abundances for latitude intervals are the sum of all relative abundances from GOS sites within that interval. (a) and (b) represent lower abundance taxa, while (c) and (d) show higher abundances. Standard deviations are indicated by gray bars and are calculated for the sites within a given interval. The '*' symbol indicates fewer than two sites within the interval and therefore no standard deviation. Connector lines are added for emphasizing the trends and do not imply continuity. (e) A global ocean net primary productivity map for 2004, with overlaying GOS sites (white dots). Productivity values are expressed as mg C m^{-2} day^{-1}.

extremes of the salinity gradient, but with higher abundance at the lower end of the gradient, supporting the brackish preference of BD1-5 clade. Effects of salinity on the metabolic capabilities of bacterioplankton have been observed previously, especially in river and estuarine systems (Bouvier & del Giorgio, 2002; Langenheder *et al.*, 2003), potentially explaining specific environment preferences. At class level, SAR202 clade lies on the lower end of the chlorophyll gradient (Fig. S3), in accordance with previous observations that this clade has abundance maximum at the lower boundary of the deep chlorophyll maximum layer (Giovannoni *et al.*, 1996). Other members of the *Chloroflexi* phylum (Keppen *et al.*, 2000) have anoxygenic photosynthesis capacity, which could be the case with the SAR202 clade. However, previous studies indicate that oxygenic and anoxygenic phototrophic bacteria co-occur in the euphotic zone (Kolber *et al.*, 2001), implying a different strategy than anoxygenic phototrophy.

Individual taxa were selected for correlation with environmental factors. Significant correlations were found only for a few groups and mainly for temperature, salinity, and chlorophyll *a* concentration values (Table S9), which were also demonstrated in the NMDS analyses. Although not all taxa are considered in this analysis, the selected taxa here are dominant members of bacterioplankton communities, and they could be the ones driving the community structure patterns observed in NMDS analyses.

Latitudinal distribution patterns were also investigated for the same selected taxa in an effort to uncover additional factors that may influence their distribution. Although some deviations were evident, the latitudinal distributions revealed two distinct types of generalized patterns; one having two abundance peaks at temperate and tropical regions (pattern 1 – Fig. 4a and c) and one having a tropical peak (pattern 2 – Fig. 4b and d). Moreover, these two patterns were observed for phylogenetically diverse groups of organisms, as well as for lower and higher abundance groups (Fig. 4 and Fig. S4). Examples of pattern 1 included surface 1 subgroup of SAR11 clade, NS4 clade of *Flavobacteria*, and SAR86 clade of *Gammaproteobacteria*, while examples of pattern 2 included *Prochlorococcus*, *Synechococcus*, SAR202 clade of *Chloroflexi*, and OM27 clade of *Deltaproteobacteria*. When comparing the latitudinal distribution of taxa in patterns 1 and 2 with global distribution of net primary production (Behrenfeld & Falkowski, 1997), we observed that pattern 1 had peaks in both high primary production areas and oligotrophic areas, while pattern 2 only peaked in oligotrophic gyre regions (Fig. 4e).

The observations for pattern 2 are consistent with previous observations for photoautotrophic bacterioplankton, where models of global distribution of *Prochlorococcus* were also found to peak in oligotrophic gyre regions (Johnson *et al.*, 2006; Follows *et al.*, 2007) (Fig. 4d). *Synechococcus* did not show lower abundances between 0- and 4-S interval (southern Pacific upwelling area), coinciding with previous observations for this organism (Rocap *et al.*, 2002) (Fig. 4d). Finally, this distribution pattern follows the general notion that these two organisms are dominant in tropical and subtropical areas. Other taxa of unknown characteristics, which fit to this pattern, may have similar metabolic capabilities, that is, photoautotrophs, with a dominant presence in tropical and subtropical regions. For example, members of SAR324 clade from the mesopelagic zone have recently been found to be capable of chemolithoautotrophy (Swan *et al.*, 2011), supporting the possibility of photoautotrophy in surface-dwelling SAR324 clade members observed in pattern 2 (Fig. 4b).

The peaks at both high primary production areas and oligotrophic areas in pattern 1 may reflect the convergent metabolic strategies exhibited by phylogenetically diverse bacterioplankton within this pattern. Autotrophic and mixotrophic functionalities, as well as implications of proteorhodopsin presence for different groups of marine bacterioplankton, have been reviewed previously (Fuhrman & Steele, 2008). Based on these reviews, we suggest that at temperate high primary production areas, these bacterioplankton exhibit a truly heterotrophic strategy and dwell on organic material produced by phytoplankton (Herndl *et al.*, 2008). At oligotrophic areas, they may depend on photoheterotrophy or mixotrophy. Because our definitions of taxa are quite broad, and there could be many subclades within all the clades considered, it will be interesting to investigate whether different subclades of certain taxa are specific to the individual locations considered in this study. This may reveal more habitat-specialized functional groups within broader clades. This will of course require a study with higher coverage in sequencing effort.

Acknowledgements

We would like to thank Mar Fernández Méndez, Petra Pjevac, and Elmar Prüsse for their assistance in phylogenetic and taxonomic analysis of the dataset. We would also like to thank Pier Luigi Buttigieg for assistance in ontological annotation of the dataset and for helpful suggestions. This study was supported by the Max Planck Society.

References

Alonso-Sáez L, Gasol JM, Lefort T, Hofer J & Sommaruga R (2006) Effect of natural sunlight on bacterial activity and differential sensitivity of natural bacterioplankton groups in northwestern Mediterranean coastal waters. *Appl Environ Microbiol* **72**: 5806–5813.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andersson AF, Riemann L & Bertilsson S (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME J* **4**: 171–181.

Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA & Thingstad F (1983) The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* **10**: 257–263.

Behrenfeld MJ & Falkowski PG (1997) Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol Oceanogr* **42**: 1–20.

Bidle KD & Azam F (1999) Accelerated dissolution of diatom silica by marine bacterial assemblages. *Nature* **397**: 508–512.

Biers EJ, Sun SL & Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the Global Ocean Sampling metagenome. *Appl Environ Microbiol* **75**: 2221–2229.

Bourne DG & Munn CB (2005) Diversity of bacteria associated with the coral *Pocillopora damicornis* from the Great Barrier Reef. *Environ Microbiol* **7**: 1162–1174.

Bouvier TC & del Giorgio PA (2002) Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. *Limnol Oceanogr* **47**: 453–470.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. *Aust J Ecol* **18**: 117–143.

Fierer N, Bradford MA & Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* **88**: 1354–1364.

Finn RD, Tate J, Mistry J *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res* **36**: D281–D288.

Follows MJ, Dutkiewicz S, Grant S & Chisholm SW (2007) Emergent biogeography of microbial communities in a Model Ocean. *Science* **315**: 1843–1846.

Fuhrman J & Hagström Å (2008) Bacterial and archaeal community structure and its patterns. *Microbial Ecology of the Oceans* (Kirchman DL, ed), pp. 45–90. Wiley-Blackwell, New York.

Fuhrman JA & Steele JA (2008) Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microb Ecol* **53**: 69–81.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV & Naeem S (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *P Natl Acad Sci USA* **103**: 13104–13109.

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL & Brown JH (2008) A latitudinal diversity gradient in planktonic marine bacteria. *P Natl Acad Sci USA* **105**: 7774–7778.

Gianoulis TA, Raes J, Patel PV *et al.* (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *P Natl Acad Sci USA* **106**: 1374–1379.

Gilbert JA, Steele JA, Caporaso JG *et al.* (2011) Defining seasonal marine microbial community dynamics. *ISME J* **6**: 298–308.

Giovannoni SJ, Rappe MS, Vergin KL & Adair NL (1996) 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *P Natl Acad Sci USA* **93**: 7979–7984.

Glöckner FO, Fuchs BM & Amann R (1999) Bacterioplankton compositions in lakes and oceans: a first comparison based on fluorescence *in situ* hybridization. *Appl Environ Microbiol* **65**: 3721–3726.

Herndl GJ, AgoguÈ H, Baltar F, Reinthaler T, Sintes E & Varela MM (2008) Regulation of aquatic microbial processes: the 'microbial loop' of the sunlit surface waters and the dark ocean dissected. *Aquat Microb Ecol* **53**: 59–68.

Hugenholtz P, Goebel B & Pace N (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* **180**: 4765–4774.

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.

Keppen OI, Tourova TP, Kuznetsov BB, Ivanovsky RN & Gorlenko VM (2000) Proposal of Oscillochloridaceae fam. nov. on the basis of a phylogenetic analysis of the filamentous anoxygenic phototrophic bacteria, and emended description of Oscillochloris and Oscillochloris trichoides in comparison with further new isolates. *Int J Syst Evol Microbiol* **50** (Pt 4): 1529–1537.

Kirchman DL, Cottrell MT & Lovejoy C (2010) The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* **12**: 1132–1143.

Kolber ZS, Plumley FG, Lang AS, Beatty JT, Blankenship RE, VanDover CL, Vetriani C, Koblizek M, Rathgeber C & Falkowski PG (2001) Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. *Science* **292**: 2492–2495.

Kostadinov I, Kottmann R, Ramette A, Waldmann J, Buttigieg PL & Glöckner FO (2011) Quantifying the effect of environment stability on the transcription factor repertoire of marine microbes. *Microb Inform Exp* **1**: 9 .

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J & Glöckner FO (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391–D395.

Kruskal J (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**: 1–27.

Langenheder S, Kisand V, Wikner J & Tranvik LJ (2003) Salinity as a structuring factor for the composition and performance of bacterioplankton degrading riverine DOC. *FEMS Microbiol Ecol* **45**: 189–202.

Larsson U & Hagström A (1979) Phytoplankton exudate release as an energy source for the growth of pelagic bacteria. *Mar Biol* **52**: 199–206.

Lauber CL, Hamady M, Knight R & Fierer N (2009) Soil pH as a predictor of soil bacterial community structure at the continental scale: a pyrosequencing-based assessment. *Appl Environ Microbiol* **75**: 5111–5120.

Lee SH & Fuhrman JA (1991) Spatial and temporal variation of natural bacterioplankton assemblages studied by total genomic DNA cross-hybridization. *Limnol Oceanogr* **36**: 1277–1287.

Lotze HK, Lenihan HS, Bourque BJ, Bradbury RH, Cooke RG, Kay MC, Kidwell SM, Kirby MX, Peterson CH & Jackson JBC (2006) Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science* **312**: 1806–1809.

Lozupone CA & Knight R (2007) Global patterns in bacterial diversity. *P Natl Acad Sci USA* **104**: 11436–11440.

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

Luo H, Benner R, Long RA & Hu J (2009) Subcellular localization of marine bacterial alkaline phosphatases. *P Natl Acad Sci USA* **106**: 21219–21223.

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

Nemergut DR, Costello EK, Hamady M *et al.* (2010) Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13**: 135–144.

Nold SC & Zwart G (1998) Patterns and governing forces in aquatic microbial communities. *Aquat Ecol* **32**: 17–35.

Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH & Wagner H (2011) *vegan: Community Ecology Package. R package version 1.17–6*. http://CRAN.R-project.org/package=vegan.

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

Pantos O, Cooney RP, Le Tissier MDA, Barer MR, O'Donnell AG & Bythell JC (2003) The bacterial ecology of a plague-like disease affecting the Caribbean coral *Montastrea annularis*. *Environ Microbiol* **5**: 370–382.

Philippot L, Bru D, Saby NPA, Čuhel J, Arrouays D, Šimek M & Hallin S (2009) Spatial patterns of bacterial taxa in nature reflect ecological traits of deep branches of the 16S rRNA bacterial tree. *Environ Microbiol* **11**: 1462–2920.

Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB & Hallin S (2010) The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* **8**: 523–529.

Pomeroy LR & Wiebe WJ (2001) Temperature and substrates as interactive limiting factors for marine heterotrophic bacteria. *Aquat Microb Ecol* **23**: 187–204.

Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, Tunlid A & Hagström Å (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**: 867–880.

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

Pruesse E, Quast C, Yilmaz P, Ludwig W, Peplies J & Glöckner FO (2011) SILVA: comprehensive databases for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches* (de Bruijn FJ, ed), pp. 393–398. John Wiley & Sons, Hoboken, NJ.

Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* **62**: 142–160.

R Development Core Team (2010) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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.

Rohwer F, Seguritan V, Azam F & Knowlton N (2002) Diversity and distribution of coral-associated bacteria. *Mar Ecol Prog Ser* **243**: 1–10.

Rusch DB, Halpern AL, Sutton G *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Seshadri R, Kravitz SA, Smarr L, Gilna P & Frazier M (2007) CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.

Swan BK, Martinez-Garcia M, Preston CM *et al.* (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the Dark Ocean. *Science* **333**: 1296–1300.

Temperton B, Gilbert JA, Quinn JP & McGrath JW (2011) Novel analysis of oceanic surface water metagenomes suggests importance of polyphosphate metabolism in oligotrophic environments. *PLoS ONE* **6**: e16499.

Virtanen R, Oksanen J, Oksanen L & Razzhivin VY (2006) Broad-scale vegetation-environment relationships in Eurasian high-latitude areas. *J Veg Sci* **17**: 519.

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

Wiebe WJ, Sheldon WM & Pomeroy LR (1992) Bacterial growth in the cold: evidence for an enhanced substrate requirement. *Appl Environ Microbiol* **58**: 359–364.

Yilmaz P, Kottmann R, Pruesse E, Quast C & Glöckner FO (2011) Analysis of 23S rRNA genes in metagenomes – a case study from the Global Ocean Sampling Expedition. *Syst Appl Microbiol* **34**: 462–469.

Yooseph S, Nealson KH, Rusch DB *et al.* (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60–66.

Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A & Ramette A (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6**: e24570.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Average *in situ* dissolved oxygen (a), nitrate (b), phosphate (c), and silicate (d) concentrations at HOT-ALOHA (2) station over a 10 year time period and depth interval between 0–5 m (solid lines), in comparison to interpolated values from the WOA05 over depth interval between 0–5 m for the same coordinates and same months (dashed lines).

Fig. S2. The relative abundances of different phyla (rows) at each GOS sampling site (columns).

Fig. S3. NMDS ordinations at phylum and class levels, with taxa weighted average scores plotted, alongside sampling sites (black dots).

Fig. S4. Latitudinal distribution patterns for other clades belonging to *Archaea*, *Actinobacteria*, *Bacteroidetes*, *Rhodobacteraceae*, SAR11, *Gammaproteobacteria*, and other *Proteobacteria*, respectively.

Table S1. Environmental variables (non-standardized) used in vector and surface fitting to NMDS ordinations.

Table S2. Bottle data extracted from HOT-ALOHA (2) station (22.75°N–158°W) between dates 1 January 2000 to 31 December 2009 and depth interval 0–5 m.

Table S3. Environment ontology term annotations of GOS sampling sites at three different levels, along with tags.

Table S4. Whole genome sequences of both marine and non-marine isolates from which the single copy domains (SCDs) were selected.

Table S5. Counts of 16S rRNA gene fragments longer than 100 bases retrieved from each GOS site.

Table S6. Number of taxa at each rank level.

Table S7. Table showing number of endemic taxa at each phylum.

Table S8. The overall taxonomic makeup of the GOS metagenome.

Table S9. Pearson rank correlation coefficients for marine clades and environmental factors.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.