Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes

Natalya Yutin, Marcelino Suzuki, Hanno Teeling, Marc Weber, J Craig Venter, Douglas B Rusch, Oded Béjà

Environmental Microbiology (2007) doi:10.1111/j.1462-2920.2007.01265.x

Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes

Natalya Yutin,[1] Marcelino T. Suzuki,[2]* Hanno Teeling,[3] Marc Weber,[3] J. Craig Venter,[4] Douglas B. Rusch[4] and Oded Béjà[1]*

[1]*Biology Department, Technion – Israel Institute of Technology, Haifa 32000, Israel.*
[2]*Chesapeake Biological Laboratory, University of Maryland Center for Environmental Sciences, PO Box 38, Solomons MD 20688, USA.*
[3]*Department of Molecular Ecology, Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 128359 Bremen, Germany.*
[4]*J. Craig Venter Institute, Rockville, MD 20850, USA.*

Summary

Aerobic anoxygenic photosynthetic bacteria (AAnP) were recently proposed to be significant contributors to global oceanic carbon and energy cycles. However, AAnP abundance, spatial distribution, diversity and potential ecological importance remain poorly understood. Here we present metagenomic data from the Global Ocean Sampling expedition indicating that AAnP diversity and abundance vary in different oceanic regions. Furthermore, we show for the first time that the composition of AAnP assemblages change between different oceanic regions with specific bacterial assemblages adapted to open ocean or coastal areas respectively. Our results support the notion that marine AAnP populations are complex and dynamic and compose an important fraction of bacterioplankton assemblages in certain oceanic areas.

Introduction

Since their rediscovery in the marine environment (Kolber *et al.*, 2000; 2001), aerobic anoxygenic photosynthetic bacteria (AAnP) were reported to exist in a variety of coastal and oceanic environments. These photoheterotrophs were detected using various techniques ranging from infrared fast-repetition-rate analysis of variable bacteriochlorophyll-*a* (BChl*a*) fluorescence (Kolber *et al.*, 2000; 2001; Koblízek *et al.*, 2005; 2006), cultivation (Allgaier *et al.*, 2003), PCR targeting of photosynthetic reaction centre genes (Béjà *et al.*, 2002; Oz *et al.*, 2005; Yutin *et al.*, 2005), real-time PCR (Schwalbach and Fuhrman, 2005; Du *et al.*, 2006), environmental genomics (Béjà *et al.*, 2002; Oz *et al.*, 2005; Waidner and Kirchman, 2005; Yutin *et al.*, 2005) and direct counts using infrared fluorescence microscopy (Schwalbach and Fuhrman, 2005; Cottrell *et al.*, 2006).

Despite these efforts, the abundance and importance of AAnPs to the flow of energy and carbon in the ocean remain poorly understood (Goericke, 2002; Schwalbach and Fuhrman, 2005; Schwalbach *et al.*, 2005). Using epifluorescence microscopy and real-time PCR, AAnPs were reported to consist of a smaller portion (up to 5%) of the total prokaryotic cells in the Pacific Ocean (Cottrell *et al.*, 2006) than originally ('at least 11%') reported (Kolber *et al.*, 2001) and to range from 2% to 16% in the Atlantic Ocean (Cottrell *et al.*, 2006; Sieracki *et al.*, 2006). Furthermore, a study by Goericke (2002) using BChl*a* measurements suggested that the contribution of BChl*a*-driven anoxygenic bacterial photosynthesis in the ocean to the conversion of light-energy is substantially lower than the previously suggested global average of 5–10% (Kolber *et al.*, 2000; 2001). These contradictory findings have at least two explanations: (i) each technique brings a certain estimation error due to inherent features, i.e. real-time PCR studies are biased due to varying binding efficiencies of the chosen primers (Yutin *et al.*, 2005), and epifluorescence microscopy is hampered by the low levels of BChl*a* in cells as well as non-specific detection of picocyanobacteria; (ii) AAnP communities are dynamic and may vary between regions and seasons.

Until now, the diversity of marine AAnP has been mainly estimated by directly amplifying *pufM* genes, encoding the M subunit of the anoxygenic photosynthetic reaction centre from environmental samples (Béjà *et al.*, 2002; Oz *et al.*, 2005; Schwalbach and Fuhrman, 2005; Yutin *et al.*, 2005; Du *et al.*, 2006), or by screening *pufM* genes in bacterial artificial chromosome (BAC) libraries (Béjà *et al.*,

Received 7 December, 2006; accepted 19 January, 2007. *For correspondence. E-mail suzuki@cbl.umces.edu; Tel. (+1) 410 326 7291; Fax (+1) 410 326 7341; E-mail beja@tx.technion.ac.il; Tel. (+972) 4829 3961; Fax (+972) 4822 5153.

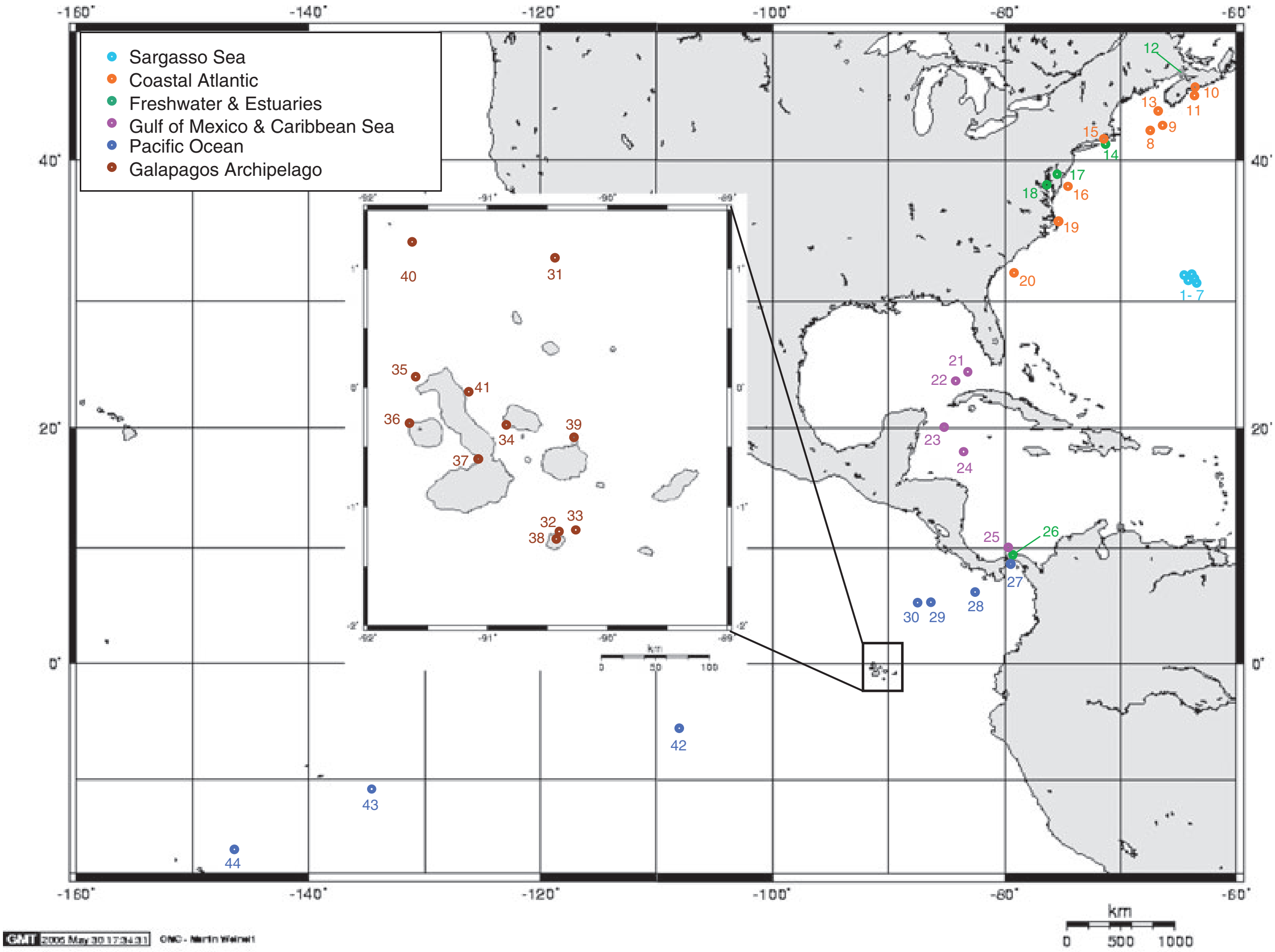

Fig. 1. The GOS transect map. The sites are numbered according to Table S1 (GOS sampling site descriptions). Different colours are used to indicate different types of environments.

2002; Oz *et al.*, 2005; Yutin *et al.*, 2005). These studies led to the discovery of novel AAnPs belonging to different groups of *Alpha*- and *Gammaproteobacteria* (Béjà *et al.*, 2002; Oz *et al.*, 2005; Yutin *et al.*, 2005). However, due to PCR- and cloning biases (Yutin *et al.*, 2005), it is difficult to estimate the proportions of different lineages within the AAnP population using these methods. In order to overcome some of the PCR-based limitations, we have used metagenomic shotgun data from the recent Global Ocean Sampling (GOS) expedition (Rusch *et al.*, 2007; Yooseph *et al.*, 2007) to characterize the distribution, composition and abundance of marine AAnPs.

Results and discussion

The GOS project produced a total of 7.6 million random sequence reads, yielding approximately six Gbp of assembled environmental DNA sequence from the North Atlantic Ocean, the Panama Canal, and East and central Pacific Ocean gyre (Fig. 1; see also Table S1 for sampling site locations and characteristics). In order to increase the coverage of particular genomes, sequence data from all sampling sites were combined for assembly (Rusch *et al.*, 2007; Yooseph *et al.*, 2007).

Using the PufM protein as a probe for anoxygenic photosynthetic bacteria (see *Experimental procedures*), 99 singletons and scaffolds (hereafter ‘scaffolds’) containing *pufM* fragments were extracted from the GOS assembled data (Table S2). The length of these scaffolds ranged from 452 to 21 305 bp and 575 sequence reads (hereafter ‘reads’) were assembled in these scaffolds.

Aerobic anoxygenic photosynthetic bacteria diversity

In order to investigate the AAnP community composition, a phylogenetic tree was reconstructed from the *pufM* sequences found on the scaffolds. The results of the *pufM* phylogenetic analysis were combined with an analysis of the *puf*-operon structure that allowed discrimination of almost all *pufM*-containing scaffolds into 12 phylogroups

(Figs 2 and 3), and also corroborated by oligonucleotide frequency analysis (see below). The distribution of the shotgun sequence data between these phylogroups is shown in Table 1. Although all reads were pooled during the assembly, no scaffolds composed of reads from both anoxic and oxic samples were obtained. Apart from scaffolds of the single anoxic station (discussed separately below), 85 scaffolds were assembled from the remaining 356 reads (Table 1). In oxic environments, only two and four reads were found belonging to phylogroups J and L respectively, and no reads were found related to phylogroups F, H. Thus, phylogroups F, H, J and L were considered as minor groups in these oxic samples.

The four most abundant phylogroups, A, B, C and D, have no cultured representatives. Moreover, these groups were almost completely missed by previous PCR-based surveys that used earlier versions of *pufM* primers (Nagashima *et al*., 1997; Achenbach *et al*., 2001; Béjà *et al*., 2002; Karr *et al*., 2003; Oz *et al*., 2005), as several base pair mismatches exist to genes in these groups. However, more recent *pufM* primers (Schwalbach and Fuhrman, 2005; Yutin *et al*., 2005) do recognize these groups. Several unusual features have been revealed from an analysis of the *puf*-operons of these phylogroups (Fig. 3); (i) the absence of *pufA* and *pufB* genes (encoding proteins for light-harvesting complex 1 that usually surround the photosynthetic reaction centre) in groups A and B. This operon organization has, so far, not been observed in any cultured organism. It is important to note here that the real AAnP capacity could not be proven based on genomics alone and more research is needed to find if these operons are indeed active; (ii) the presence of the *pufX* gene in the *puf*-operons of bacteria from group A. Although until recently, the presence of the *pufX* gene had only been reported for the anaerobic *Rhodobacter* lineage, evidences of *pufX* have been reported for other uncultured bacteria from oxic estuarine (Waidner and Kirchman, 2005) and marine (Yutin and Béjà, 2005) environments. Based on our data set, we propose that PufX-containing reaction centres are common among marine AAnP bacteria, as it was present in 35 of the 85 AAnP scaffolds described; (iii) group D bacteria have a unique order of genes encoding the reaction centre core proteins because all three subunits of the reaction centre are colocated within the same operon (*pufLMH*).

Based on reads proportions within a given station (Fig. 4), phylogroups A, B, C and D were estimated to be prevalent in the majority of pelagic AAnP communities (Sargasso Sea samples and almost all Pacific samples, including those off the Galapagos Islands; Fig. 4). As these groups were probably missed by most PCR-based AAnP diversity studies, the significance of AAnP bacteria in open ocean microbial communities might have been underestimated so far.

Phylogroup E contains the *pufX* gene and was mainly found in several stations at the North-western Atlantic coast. There is, however, a BAC clone (EBAC60D04) with 99% DNA identity to one of the phylogroup E scaffolds that has been obtained from the Pacific Ocean (Béjà *et al*., 2002).

Phylogroup G represents *Roseobacter*-like bacteria. One of the scaffolds from this group is identical to the Red Sea (Oz *et al*., 2005) environmental BAC clone eBACred25D05, and another one is identical to the Mediterranean Sea BAC clone BACmed 31B01 (unpublished data). Members of the *Roseobacter* lineage are well 1 known to be represented across a variety of marine habitats (Buchan *et al*., 2005) and our observations support this notion as group G bacteria were mainly observed at coastal and open water stations in the Pacific as well as the Atlantic Ocean.

The closest cultured relatives of group I bacteria are *Betaproteobacteria* HTCC528 from a freshwater lake (Page *et al*., 2004), *Rhodoferax antarcticus* from an Antarctic microbial mat (Madigan *et al*., 2000), and *Rhodoferax fermentans* isolated from a sewage ditch (Hiraishi *et al*., 1991). Based on this high similarity, phylogroup I likely represents *Betaproteobacteria* scaffolds. *Rhodoferax*-related bacteria are found widely distributed in freshwater environments (Glockner *et al*., 2000; Page *et al*., 2004) and accordingly, GOS sequences belonging to this group were only detected in estuarine or freshwater samples (Fig. 4). One estuarine fosmid clone [DelRiverFos06H03 (Waidner and Kirchman, 2005)] is also affiliated with these sequences.

Phylogroup K contains *Gammaproteobacteria* representatives [*Congregibacter litoralis* KT71 (Eilers *et al*., 2001) and BAC clones EBAC65D09 and EBAC29C02 (Béjà *et al*., 2002)], all related to the OM60 clade (Rappé *et al*., 1997). Despite their close relationships based on *pufM* phylogeny, AAnPs in this group possess two different types of *puf*-operon structures: some have *pufC* downstream of *pufM* while others have *pufA* and *pufB* (Fig. 3) indicating that this group might in fact be polyphyletic. Members of group K were found at the North-western Atlantic coast and at some stations off the Galapagos Islands.

Anaerobic photosynthetic scaffolds assembled from an anoxic sample

The single anoxic sample in this set of the GOS expedition (hypersaline lagoon, Galapagos; dissolved oxygen, 0.06 mg l^{-1}), contributed 219 reads to our data set (Table 1). Half or these reads were assembled in a 21 255-bp-long scaffold (#1096627135419) belonging to group H (Fig. 2). This group represented the most abundant anoxygenic phototrophs at this station, and currently contains no known cultured relatives. The *puf*-operon

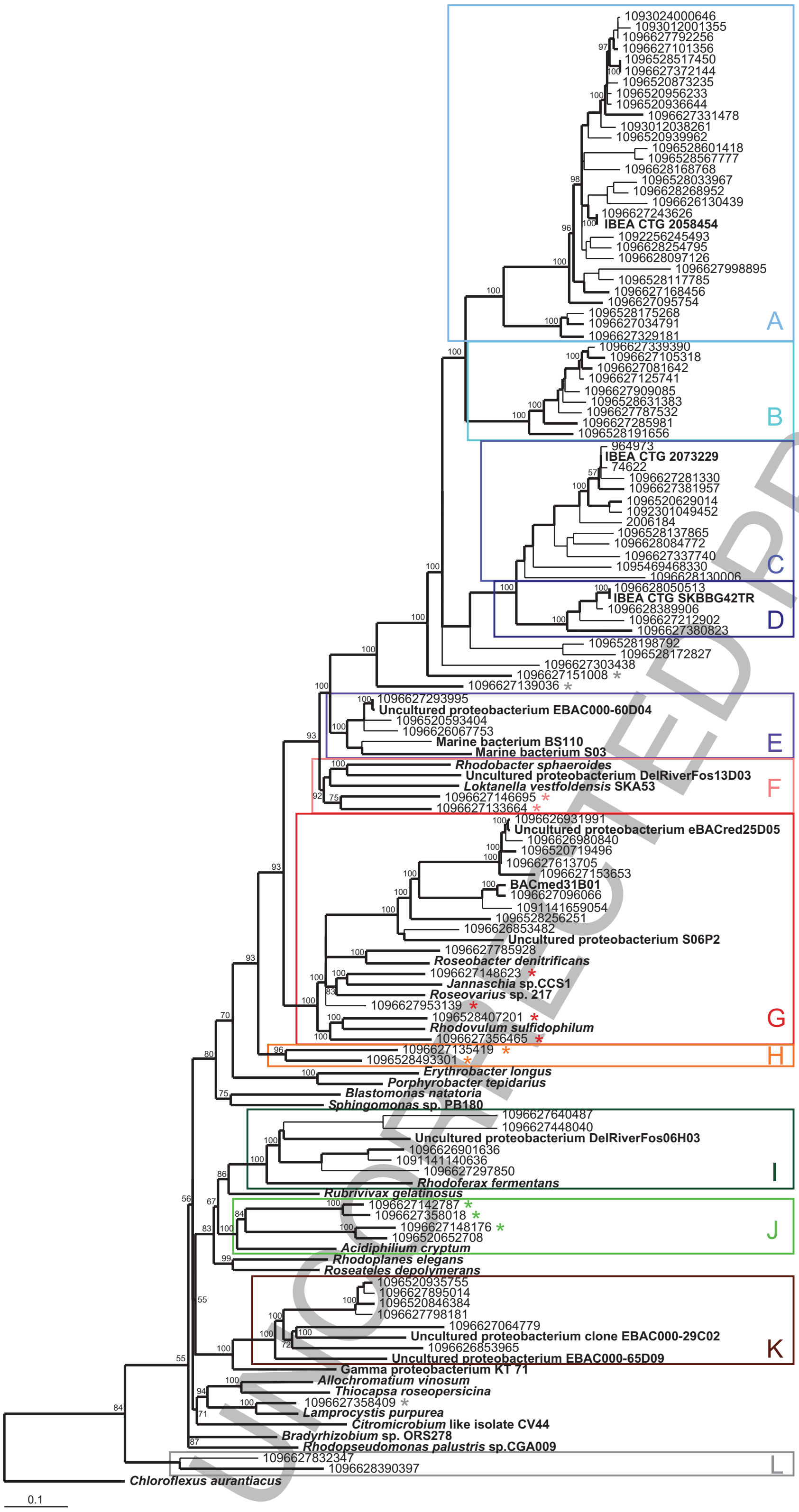

Fig. 2. *pufM* phylogenetic tree. The tree is based on a Bayesian tree to which short sequences were added by ARB_PARSIMONY. Reference sequences retrieved from the GenBank are marked in bold. IBEA_CTG clones belong to the Sargasso Sea project (Venter *et al*., 2004) whose sequence data were included in GOS assembly. Branches found on the initial Bayesian tree are shown in bold lines. The numbers on nodes represent branch confidence values. Asterisks indicate scaffolds assembled by reads from the anoxic sample. Coloured boxes mark the 12 phylogroups defined in this study.

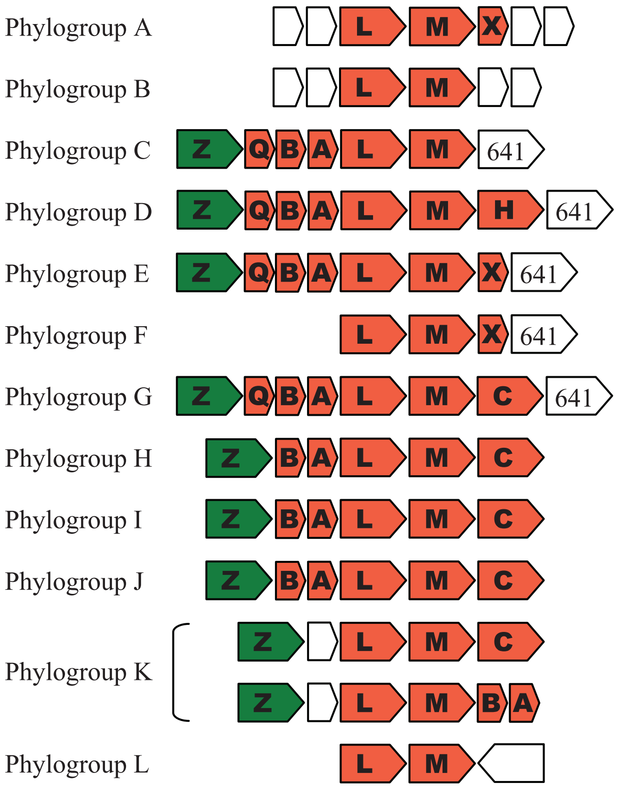

Fig. 3. Diversity of photosynthetic operonal organization revealed in the GOS data set. *puf*-operons are shown in red. Z, chlorophyllide reductase Z subunit, is marked in green. 641, deoxyxylulose-5-phosphate synthase (*orf* 641). Other genes are labelled in white.

structure of this scaffold is *pufBALMC* (Fig. 3). Scaffold number 1096627139036 represents the second most abundant bacterium at this station and is somewhat related to the *Rhodobacter* clade with a *puf*-operon structure identical to that of *Rhodobacter* species: *pufQBALMX*. Currently no cultured strains closely related to this phylotype exist. Other abundant *pufM*-containing scaffolds found at station 38 were related to *Loktanella vestfoldensis* (phylogroup F), *Rhodovulum sulfidophilum* (phylogroup G), Antarctic clones LFc1 and LFc15 (Karr *et al*., 2003) (phylogroup J; the sequences from these clones were short and are not shown in Fig. 2), and *Lamprocystis purpurea* (scaffold 1096627358409). Not surprisingly, anoxygenic phototrophs at this station were quite unique as no anoxygenic phototrophs found at this anoxic sample were observed in any of the oxic samples.

The assembly efficiency at this sample was remarkably higher than in the oxic samples. Four nearly completed genomes and the highest ratio of unassembled/assembled in >10 000-bp-long scaffolds reads were found here (Rusch *et al*., 2007) and no singletons coming from station 38 were observed in the represented data set (Table S1). This success of assembly may be attributed to lower species richness at this station and/or to the large sequencing effort performed at this station (694 642 reads).

Oligonucleotide frequency analysis of pufM*-containing scaffolds from the GOS data set*

The validity of the phylogroups defined on *pufM* phylogenetic reconstruction and operon organization was further evaluated by genomic signature analysis. As comparative genomics has revealed in recent years, the frequencies of short oligonucleotides in genomes act like a species-specific fingerprint and furthermore carry a weak phylogenetic signal (Pride *et al*., 2003). While the analysis of such genomic signatures cannot compete with gene-based phylogenetic reconstruction, regarding the sophistication of the underlying mathematics and thus resolution, it has the advantage of not being limited to genes. Instead, the entire DNA sequence can be analysed which is particularly interesting in metagenomics,

Table 1. Summary of *pufM*-associated data extracted from the GOS data set.

	Total		Oxic		Anoxic	
Phylogroup	Reads	Scaffolds	Reads	Scaffolds	Reads	Scaffolds
A	78	30	78	30	0	0
B	47	9	47	9	0	0
C	36	12	36	12	0	0
D	29	4	29	4	0	0
E	61	3	61	3	0	0
F	9	2	0	0	9	2
G	63	14	53	10	10	4
H	107	2	0	0	107	2
I	14	5	14	5	0	0
J	18	4	2	1	16	3
K	26	6	26	6	0	0
L	4	2	4	2	0	0
Others	83	6	6	3	77	3
Sum	575	99	356	85	219	14

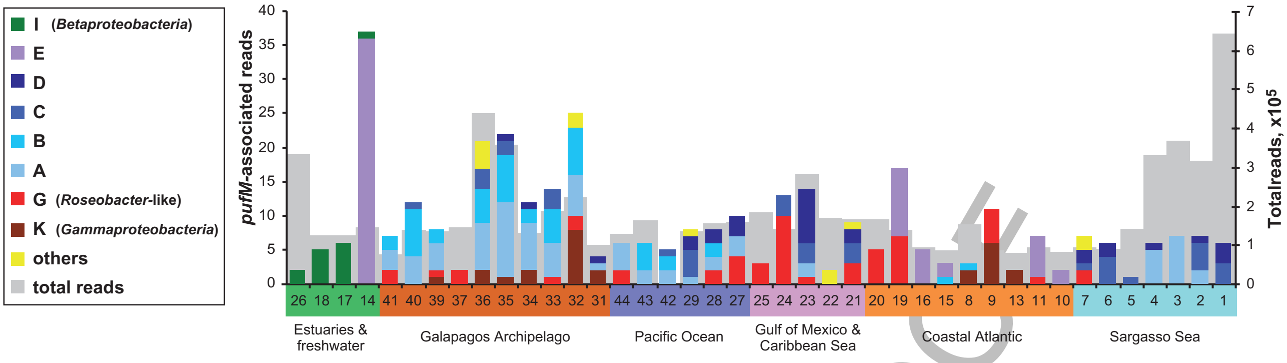

Fig. 4. AAnP population compositions along the GOS transect. Colours used to represent different types of environments are the same as in Fig. 1, and colours representing the eight major phylogroups are according to those used in Fig. 2. *pufM*-associated reads are reads included in *pufM*-containing scaffolds. Note that samples 5, 6 and 7 are different size fractions from the same station: sample 5, 20–3 μm; sample 6, 3–0.8 μm; sample 7, 0.8–0.1 μm.

where frequently only short fragments or only partial genes are obtained. Due to the shortness of many of the scaffolds in this study, only di- and trinucleotide frequencies could be used because statistics on longer oligonucleotides require longer sequences. While this did not allow computing well-separated coherent *de novo* clusters, a cluster analysis of the data revealed striking congruities with the *pufM* phylogenetic reconstruction (Fig. 5). Sixty-one per cent of the scaffolds formed clusters uniquely represented by sequences from distinct phylogroups assigned by phylogenetic and operon analysis.

Aerobic anoxygenic photosynthetic bacteria biogeography

Although the occurrence of specific gene-anchored reads and scaffolds in shotgun sequence data is a likelihood event (especially for a rare gene as *pufM*), our analysis of AAnP community composition at different stations reveal some interesting trends (Fig. 4). The *Roseobacter*-related phylogroup (G) appears to be the most ubiquitous across different environments. In addition, phylogroup G constitutes a significant part of the AAnP communities at mesotrophic stations (stations 24–28 surrounding the Panama Canal; coral reef atoll at station 44; some coastal stations and some stations off the Galapagos Islands) and a minor part of oligotrophic AAnP communities (Pacific Ocean Gyre and Sargasso Sea stations). Phylogroups A and B, not previously described, are the main AAnPs in oligotrophic regions. Phylogroups E and K subsume coastal species, whereas phylogroups C and D represent mostly offshore species. Group I bacteria, attributed to a freshwater *Betaproteobacteria* clade, also composed a significant part of AAnPs in all estuarine samples, indicating that they might thrive in saline environments (the salinity at the station 14 was 26.5 p.p.t.).

Aerobic anoxygenic photosynthetic bacteria abundance estimation

While the GOS expedition was originally intended to the discovery of new genes and organisms, and evaluation of microbial genetic diversity (Rusch *et al*., 2007), our study was aimed to survey AAnP community compositions and to gain insights on specific AAnP abundances along the GOS transect. An important question in the study of AAnPs is what per cent of total bacteria (or total microbes) in the community AAnPs comprise. We inferred AAnP relative abundances by the relative abundances of anoxygenic photosynthetic genes in these samples using different metrics. In all anoxygenic phototrophs genomes reported to date, the *pufM* gene was found as a single-copy gene and was therefore used to estimate AAnP numbers in our samples. Besides the marker for AAnP bacteria, we used the *recA* gene coding a critical DNA repair enzyme and considered to be a single-copy gene present in all bacterial genomes; thus, it is a suitable estimator of total bacterial genomes in the sample (Venter *et al*., 2004; Howard *et al*., 2006). As the length of *pufM* and *recA* genes is, on average, similar, initially we assumed that the ratio between the number of *pufM* and *recA* reads reflected the ratio between AAnP and total bacteria in our samples (data not shown).

These *pufM* and *recA* reads are meant as sequence reads strictly containing at least a part of the *pufM* or the *recA* genes. However, in addition, our data set was composed of scaffolds assembled from different reads from the entire data set, some containing fragments of other genes situated in certain proximity to the *pufM* or the *recA* genes. This information was used as a second metric of relative abundance, as common genomes should result in longer scaffolds. All reads (and not just those containing *pufM*) assembled in *pufM*-containing scaffolds were counted as '*pufM*-associated' reads in Fig. 4. While in

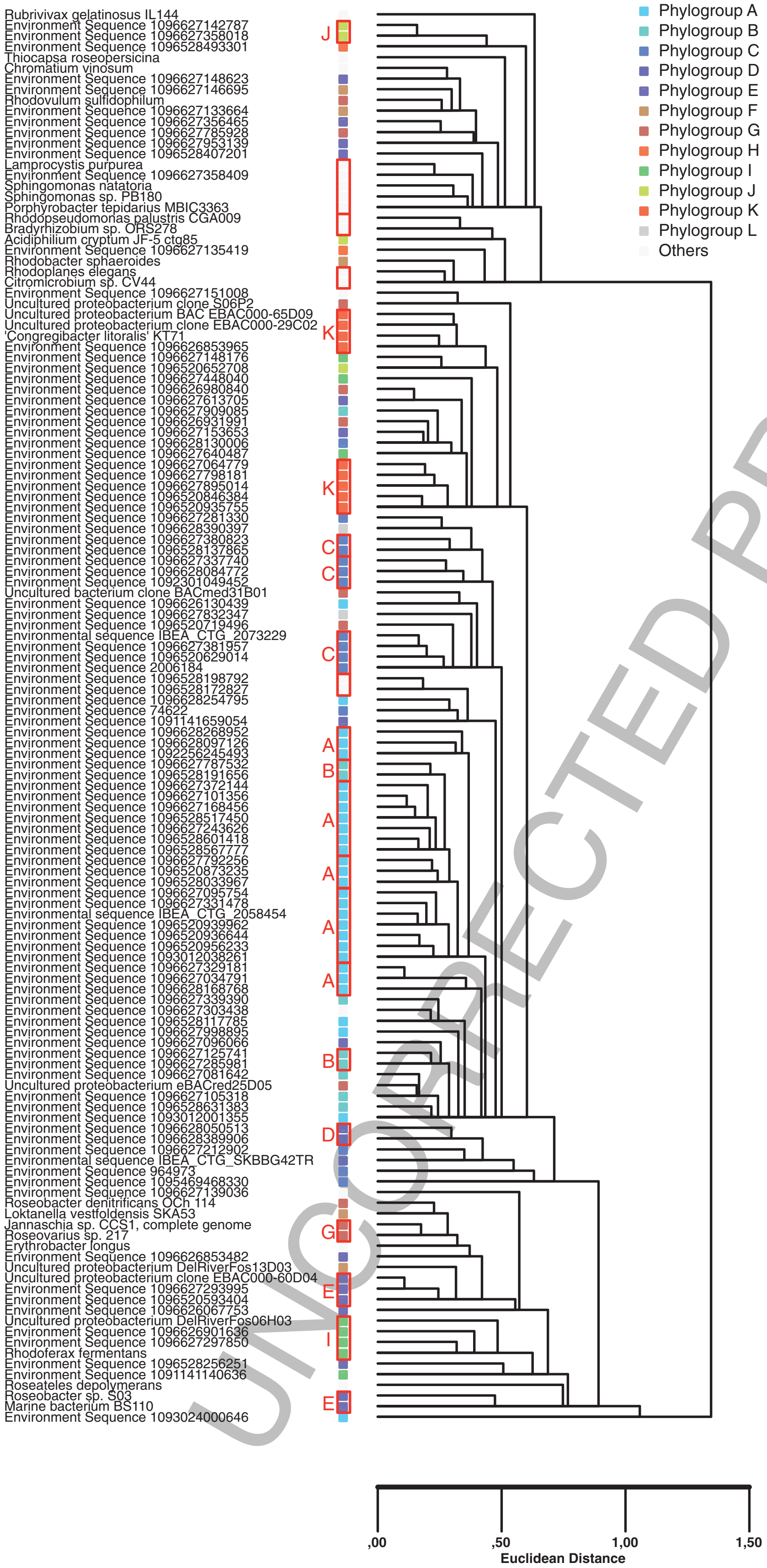

Fig. 5. Hierarchical cluster analysis of the sequences' di- and trinucleotide frequencies. Different phylogroups are indicated by distinct colours. Clusters consistent with the phylogenetic analysis are grouped by red rectangles. See *Experimental procedures* for further details.

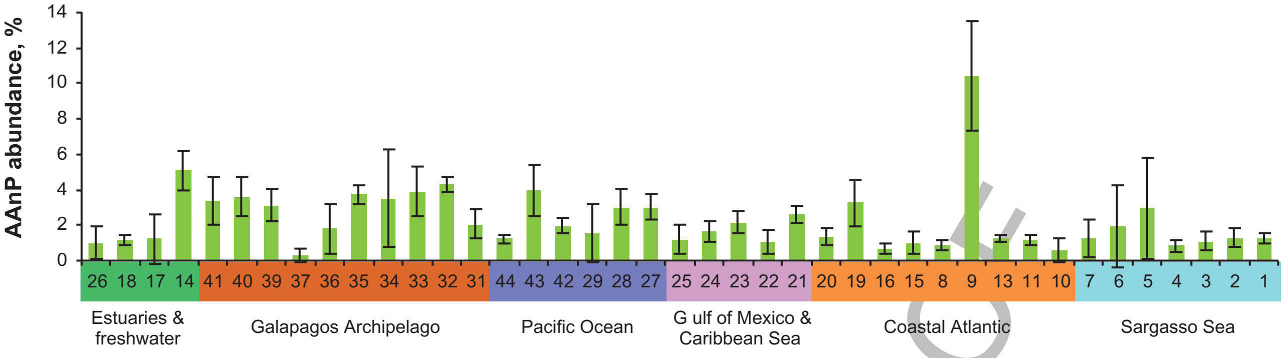

Fig. 6. Estimated AAnP abundances along the GOS transect.

poorly assembled scaffolds (composed of one to two reads, and comprising 69% of *pufM*-containing scaffolds) the number of *pufM* reads and *pufM*-associated reads are nearly equal, this is not the case in large scaffolds. For instance, the high number of *pufM*-associated reads at station 14 (Fig. 4) is the result of the assembly of a 17 830-bp-long scaffold, containing 36 (out of 57) reads belonging to this station (Table S2). Again, as abundant genomes should result in longer scaffolds, it was not surprising that 55% of *recA*-associated reads in the entire data set were assembled in only 16 of 5104 scaffolds, ranging from 1 to 10 723 reads.

A third metric was also developed as measure for approximating the total number of *pufM* (or *recA*) reads per station based on the assembled data. This measure, termed 'read equivalents' (see *Experimental procedures* for details), is not depended (at least, not directly) on the scaffold size, and, thus, is an estimate of total *pufM* reads in each of the individual stations. The read equivalent measure was developed to approximate the number of reads targeting certain genes from scaffold assembly data. It is based on the number of reads participating in a given scaffold (scaffold size), and on the ratio of scaffold length/gene length. Read equivalents are more suitable for longer scaffolds, where statistic approaches and averaging-like coverage concept are applicable. In our case, 'the end effects' are considerable (only a part of a *pufM* gene appears on the scaffold). Based on the calculations, short *pufM* end-sequences in long scaffolds underestimate read equivalents, while longer *pufM* end-sequences in short scaffolds overestimates the measure. Assuming the proportion of these type of sequences are similar for different genes (i.e. *recA*), the normalization step should compensate for these errors.

Whereas 5104 scaffolds containing *recA* genes were found in the entire data set, only 99 scaffolds were found containing *pufM*. Because of this low number of scaffolds, we increased the reliability of our estimations by adding to the analysis two additional AAnP unique markers, namely the *pufL* gene coding for the L subunit of the anoxygenic photosynthetic reaction centre and the *bchX* gene coding for the X subunit of chlorophyllide reductase in the bacteriochlorophyll biosynthetic pathway. Both genes have roughly the same length as *pufM* and *recA* and, as can be seen in Fig. S1, there was a very high correlation between these three AAnP markers in most stations measured by the number of gene-associated reads, as well as read equivalents. For each sampling station, *pufM*, *pufL* and *bchX* read equivalents were calculated, normalized by *recA* read equivalents and averaged to produce the AAnP abundance evaluation shown in Fig. 6. Aerobic anoxygenic photosynthetic bacteria abundance varied between different marine stations. The highest AAnP percentages were observed at a coastal (10%; station 9, Browns Bank, Gulf of Maine) and an estuarine (5%; station 14, Newport Harbor) station. In several Pacific Ocean samples, the AAnP abundance was near 4% (see open ocean station 43 and Galapagos stations 32–35). These estimates therefore imply that AAnPs are found in relatively similar proportions in both oligotrophic and eutrophic environments and are important component of the total upper ocean microbial community.

Potential biases of sampling and metagenomics data set processing

The results reported here are for samples in the 0.1–0.8 μm size range. Aerobic anoxygenic photosynthetic bacteria have been reported to be larger than the average bacterioplankton (Sieracki *et al.*, 2006) and in addition, some AAnPs might be associated with larger particles (as symbionts of eukaryotic cells, or attached to marine snow) and might produce chains as reported for *Erythrobacter* (Yurkov and Beatty, 1998); see differences in population composition in samples 5, 6 and 7 in Fig. 4, which are different size fractions of the same station). Thus there is a possibility that our estimates of AAnP diversity and relative abundance are underestimates. On the other hand, high cell densities in some water samples may have clogged 0.8 μm filters producing a bias towards smaller cells. Finally, besides biases introduced by sampling methods, many known biases might result from differential DNA extraction, cloning using *E. coli* as a host, and other common problems associated with metagenomic assembly like chimera formation and, intraspecies sequence variability.

In addition, pooling the raw sequencing results prior to assembly, applied for the first time in this data treatment, raised a new challenging statistical problem. In single-station metagenomic assembly, scaffold length and coverage are functions of the amount of sequencing performed at the station, of the species richness, and the length and the abundance of particular genomes contributing to the scaffolds. When scaffolds are assembled from reads coming from different sampling sites, lengths and coverages depend on the same parameters, but from all sampling sites combined and thus high abundance of a genome at one of the sampling sites increased the chance of its assembly at all sampling sites. In our case, this meant an increased probability to find the same AAnPs across sampling sites. Fortunately, this did not appear to be case in our study as 72 from the 99 *pufM*-containing scaffolds were assembled from reads from single stations. However, this issue remains to be resolved for larger future metagenomic data sets.

Aerobic anoxygenic photosynthetic bacteria relative abundance values calculated in this study are based on approximate relative abundances of *pufM*-*pufL*-*bchX*-containing bacteria, with the assumption that these genes, as well as the *recA* gene, are single-copy in their genomes. In most anoxygenic phototrophs genomes reported to date, including the genome of the AAnP *Roseobacter denitrificans* Och 114 (Swingley *et al*., 2007), the *pufM* gene was found as a single-copy gene and was therefore used to estimate AAnP numbers in our samples. However, in two reported cases, *Roseobacter litoralis* and *Staleya guttiformis*, the *pufM* gene was also found on an extrachromosomal linear plasmids (Pradella *et al*., 2004). The number of copy of these linear plasmids per cell was not determined and currently it is not known how general this phenomenon is.

The *recA* gene was used to normalize AAnP abundance estimates because these genes are the most often employed single-copy gene normalizers (Venter *et al*., 2004; Howard *et al*., 2006). However, other genes than the *recA* gene may be used for the normalization. We checked how our estimates were influenced when alternate 'single-copy protein' genes like *rpoB* (RNA polymerase B) and *gyrA* (DNA gyrase subunit A) previously used to estimate the number of genomes represented in metagenomic libraries (Venter *et al*., 2004). Aerobic anoxygenic photosynthetic bacteria abundances normalized by *rpoB* and *gyrA* genes were calculated and compared with those obtained with the *recA* gene (Fig. S2). The calculations were the same as for *recA*, with an additional gene length normalization step, because *rpoB* and *gyrA* genes both are significantly longer than the *c.* 1000 bp of *rec*A (*c.* 3600 and 2700 bp respectively). For most stations, values obtained were similar (within the standard deviations calculated from *pufM-pufL-bchX* averages using *recA* as normalizer).

As the algorithm for searching *recA*-containing contigs does not discriminate bacterial RecAs from eukaryotic and archaeal homologues (RAD family proteins), the abundances obtained represent per cents of total cells in the community rather than percents of total bacteria. However, due to the size fractionation we very likely excluded the vast majority of *Eukarya* from the samples, and thus these numbers are comparable to the percentages of AAnPs measured relative to total (DAPI) bacterial counts. The *rpoB* (Walsh *et al*., 2004; Case *et al*., 2007) and *gyrA* (Guipaud *et al*., 1997; Wall *et al*., 2004) genes may also have close analogues in archaea and plastid DNA.

Although metagenomics has been so far mostly used for culture- and PCR-independent gene discovery, recent studies have used metagenomic assembly for prediction of viral community diversity and species richness (Angly *et al*., 2006). We believe that our preliminary calculation raises the importance of refined models combining bacterial population structure and diversity parameters using metagenomic data.

Concluding remarks

Combining the GOS abundance results with the results accumulated using BChl*a*-based biophysical measurements (Kolber *et al*., 2000; 2001), real-time PCR and infrared fluorescence microscopy (Schwalbach and Fuhrman, 2005; Cottrell *et al*., 2006; Du *et al*., 2006), we suggest that AAnP loads vary significantly between different regions and represent a dynamic component of marine bacterioplankton. Furthermore, our results show that not only abundance but also AAnP composition varies between different oceanic regions. This is, to our knowledge, the first time that the AAnP population composition is estimated in a global biogeographical context.

Experimental procedures

Global Ocean Sampling sample collection, shotgun cloning, primary assembly and extraction of AAnP-related data

Water samples were collected from February 2003 to May 2004 along a North-South transect between 45°N in the Atlantic Ocean and 15°S in the Pacific Ocean. Samples were collected from a wide range of habitat types, including oceanic and coastal seawater, freshwater and hypersaline lakes, estuaries, and areas surrounding oceanic islands. The coordinates and different characteristics of sampling stations are presented in Table S1. Sampling procedures, library construction, shotgun sequencing and assembly are described in Rusch and colleagues (2007). Shotgun data used in this study were obtained mainly from 0.8 to 0.1 μm sized planktonic fractions (see Table S1 for sample fraction sizes). The assembly was performed on entire pool of GOS sequence reads, with an overlap cut-off of 98% identity, whereas minimal length of an overlap was 40 bp.

Scaffolds related to anoxygenic bacteria were extracted by a sequence recently reported similarity clustering approach (Yooseph *et al.*, 2007). Briefly, protein sequences produced from the assembled scaffolds were clustered with a non-redundant set of publicly available sequences within the NCBI-nr, NCBI Prokaryotic Genomes, TIGR Gene Indices and Ensemble data sets based on pair-wise sequence similarity. Clustering was based on full-length sequences, rather than domains, and incorporated length-based thresholds to address fragmentary sequences thereby minimizing the clustering of unrelated proteins. In this way, 99 *pufM*-containing scaffolds were revealed. Contributions of each sampling station to every scaffold are given in Table S2. Using the methodology, 109 *pufL*, 109 *bchX*, 5104 *recA*, 8392 *gyrA* and 10 482 *rpoB* gene-containing scaffolds were detected in this data set.

Phylogenetic tree reconstruction

pufM phylogenetic analysis was initially performed using 36 reference sequences from cultured species and environmental genomic clones retrieved from GenBank and 46 GOS scaffolds containing significant sequence overlap (*c.* 750 bp) with the reference sequences [positions homologous to positions 7–736 of the *Rhodobacter sphaeroides* sequence (AJ010302)]. Using ARB (Ludwig *et al.*, 2004), the GOS scaffold nucleotide sequences were imported into a previously described *pufM* database (Yutin *et al.*, 2005), translated into amino acids and aligned. Thereafter, the alignment was manually corrected. The resulting protein alignment was used to realign (back-translate) nucleotide sequences in ARB, and this nucleotide alignment was used in all subsequent phylogenetic analyses. Aligned nucleotide sequences were exported using a filter that excluded positions where gaps outnumbered characters, and kept the nucleotides in frame (720 total nucleotide positions). From this filtered alignment a phylogenetic tree was reconstructed by Bayesian inference using the MrBayes 3.0 program (Ronquist and Huelsenbeck, 2003) with the General Time Reversible model and rates varying according to codon positions. Four parallel chains of 1 million generations were run, trees were sampled every 100 generations, and 600 'burnin' trees were excluded from the consensus tree. This consensus tree was imported into ARB and 53 shorter nucleotide sequences were aligned as above and added to the Bayesian tree using the ADD_BY_PARSIMONY algorithm and the same filter.

Calculating AAnP abundances along the GOS transect

pufM read equivalents at all sampling sites were calculated as described below.

For each (*i*th) scaffold, its coverage is expressed as:

$$a_i = N_i \cdot L/S_i \qquad (1)$$

where N_i is the total number of reads in the *i*th scaffold, *L* is the mean read length (842 bp) in the entire data set, and S_i is the *i*th scaffold length.

The total number of bp associated with the *pufM* sequence at the *i*th scaffold was defined as:

$$B_i = g_i \cdot a_i \qquad (2)^*$$ 3

where g_i is the length of the *pufM* fragment on the *i*th scaffold.

The contribution of *j*th station reads to the *i*th scaffold:

$$m_{ij} = n_{ij}/N_i \qquad (3)$$

where n_{ij} is the number of reads from the *j*th station participating in the *i*th scaffold; $N_i = \Sigma^j n_{ij}$.

For each (*j*th) station, the number of bp associated with a given (*i*th) *pufM* fragment is calculated as:

$$b_{ij} = B_i \cdot m_{ij} \qquad (4)$$

pufM read equivalents (*r*) were defined as:

$$r_{ij} = b_{ij}/L \qquad (5)$$

The read equivalents approximate the number of reads containing *i*th *pufM* fragment at the *j*th station. From Eqs 1–5, $r_{ij} = n_{ij} \cdot g_i/S_i$.

The total number of *pufM* read equivalents at the *j*th station:

$$R_j = \sum^i r_{ij}$$

An example of *pufM* read equivalent calculation for one of the sampling sites is shown in *Supplementary materials*.

Total numbers of read equivalents for *recA* and all other normalizer genes were calculated. The *pufM*-based AAnP abundance at each station was estimated as:

$$A_j^{pufM} = \frac{R_j^{pufM}}{R_j^{recA}} \cdot 100\%$$

Additionally, scaffolds containing fragments of *pufL* and *bchX* genes were extracted from the GOS data set. *pufL* and *bchX* read equivalents were calculated; *pufL* and *bchX*-based AAnP abundances at each station were estimated as: $A_j^{pufL} = \frac{R_j^{pufL}}{R_j^{recA}} \cdot 100\%$; $A_j^{bchX} = \frac{R_j^{bchX}}{R_j^{recA}} \cdot 100\%$.

Genome signature analysis

The DNA sequences of the 99 scaffolds and the 36 reference sequences used in this study were imported into TETRA (Teeling *et al.*, 2004) where four length-independent parameters were computed: relative dinucleotide and trinucleotide frequencies, dinucleotide relative abundances and Markov model-based trinucleotide *z*-scores. These data were exported and imported into Aabel (Gigawiz), where an unweighted hierarchical cluster analysis with the Euclidian distance as distance measure was computed for all 160 data columns.

Acknowledgements

We would like to thank B. Fuchs for providing unpublished genome sequences and M. Shmoish for fruitful discussions. This work was supported in part by grant from the Israel Science Foundation, a grant from the Israeli Ministry of Science and Technology and an EMBO YIP award (O.B.), NSF Grant OCE-0550547 (M.T.S.) and by the U.S. Department of Energy Office of Science's Biological and Environmental Research Program grant and the Gordon and Betty Moore Foundation (J. Craig Venter Institute). We thank the Governments of Canada, Mexico, Honduras, Costa Rica, Panama, Ecuador and French Polynesia/France for facilitating sampling activities. All sequencing data collected from waters of the above named countries remain part of the genetic patrimony of the country from which they were obtained.

References

Achenbach, L.A., Carey, J., and Madigan, M.T. (2001) Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Appl Environ Microbiol* **67:** 2922–2926.

Allgaier, M., Uphoff, H., and Wagner-Dobler, I. (2003) Aerobic anoxygenic photosynthesis in *Roseobacter* clade bacteria from diverse marine habitats. *Appl Environ Microbiol* **69:** 5051–5059.

Angly, F., Felts, B., Breitbart, M., Salamon, P., Edwards, R., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., *et al.* (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415:** 630–633.

Buchan, A., González, J.M., and Moran, M.A. (2005) Overview of the marine *Roseobacter* lineage. *Appl Env Microbiol* **71:** 5665–5677.

Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F., and Kjelleberg, S. (2007) The 16S rRNA and *rpoB* genes as molecular markers for microbial ecology. *Appl Environ Microbiol* (in press). 4

Cottrell, M.T., Mannino, A., and Kirchman, D.L. (2006) Aerobic anoxygenic phototrophic bacteria in the Mid-Atlantic Bight and the North Pacific Gyre. *Appl Environ Microbiol* **72:** 557–564.

Du, H., Jiao, N., Hu, Y., and Zeng, Y. (2006) Real-time PCR for quantification of aerobic anoxygenic phototrophic bacteria based on *pufM* gene in marine environment. *J Exp Mar Biol Ecol* **329:** 113–121.

Eilers, H., Pernthaler, J., Peplies, J., Glockner, F.O., Gerdts, G., and Amann, R. (2001) Isolation of novel pelagic bacteria from the german bight and their seasonal contributions to surface picoplankton. *Appl Environ Microbiol* **67:** 5134–5142.

Glockner, F.O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., and Amann, R. (2000) Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Appl Env Microbiol* **66:** 5053–5065.

Goericke, R. (2002) Bacteriochlorophyll *a* in the ocean: is anoxygenic bacterial photosynthesis important? *Limnol Oceanogr* **47:** 290–295.

Guipaud, O., Marguet, E., Noll, K.M., de la Tour, C.B., and Forterre, P. (1997) Both DNA gyrase and reverse gyrase are present in the hyperthermophilic bacterium *Thermotoga maritima*. *Proc Natl Acad Sci USA* **94:** 10606–10611.

Hiraishi, A., Hoshimao, Y., and Satoh, T. (1991) *Rhodoferax fermentans* gen. nov. and sp. nov., a phototrophic purple non-sulfur bacterium previously referred to as the '*Rhodocyclus gelatinosus*-like' group. *Arch Microbiol* **153:** 330–336.

Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Burgmann, H., Welsh, R., *et al.* (2006) Bacterial taxa that limit sulfur flux from the ocean. *Science* **314:** 649–652.

Karr, E.A., Sattley, W.M., Jung, D.O., Madigan, M.T., and Achenbach, L.A. (2003) Remarkable diversity of phototrophic purple bacteria in a permanently frozen Antarctic lake. *Appl Environ Microbiol* **69:** 4910–4914.

Koblízek, M., Ston-Egiert, J., Sagan, S., and Kolber, Z.S. (2005) Diel changes in bacteriochlorophyll *a* concentration suggest rapid bacterioplankton cycling in the Baltic Sea. *FEMS Microbiol Ecol* **51:** 353–361.

Koblízek, M., Falkowski, P.G., and Kolber, Z.S. (2006) Diversity and distribution of anoxygenic phototrophs in the Black Sea. *Deep Sea Res II* (in press). 5

Kolber, Z.S., Van Dover, C.L., Niderman, R.A., and Falkowski, P.G. (2000) Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407:** 177–179.

Kolber, Z.S., Plumley, F.G., Lang, A.S., Beatty, J.T., Blankenship, R.E., VanDover, C.L., *et al.* (2001) Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. *Science* **292:** 2492–2495.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, et al. (2004) ARB: a software environment for sequence data. *Nucleic Acid Res* **32:** 1363–1371. 6

Madigan, M.T., Jung, D.O., Woese, C.R., and Achenbach, L.A. (2000) *Rhodoferax antarcticus* sp. nov., a moderately psychrophilic purple nonsulfur bacterium isolated from an Antarctic microbial mat. *Arch Microbiol* **173:** 269–277.

Nagashima, K.V.P., Hiraishi, A., Shimada, K., and Matsuura, K. (1997) Horizontal transfer of genes coding for the photosynthetic reaction centers of purple bacteria. *J Mol Evol* **45:** 131–136.

Oz, A., Sabehi, G., Koblízek, M., Massana, R., and Béjà, O. (2005) *Roseobacter*-like bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. *Appl Environ Microbiol* **71:** 344–353.

Page, K.A., Connon, S.A., and Giovannoni, S.J. (2004) Representative freshwater bacterioplankton isolated from Crater Lake, Oregon. *Appl Env Microbiol* **70:** 6542–6550.

Pradella, S., Allgaier, M., Hoch, C., Pauker, O., Stackebrandt, E., and Wagner-Dobler, I. (2004) Genome organization and localization of the *pufLM* genes of the photosynthesis reaction center in phylogenetically diverse marine Alphaproteobacteria. *Appl Environ Microbiol* **70:** 3360–3369.

Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13:** 145–158.

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

Ronquist, F., and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.

Rusch, D.B., Halpern, A.L., Heidelberg, K.B., Sutton, G., Williamson, S.J., Yooseph, S., *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: I, The northwest Atlantic through the eastern tropical Pacific. *PLoS Biol* (submitted). 7

Schwalbach, M.S., and Fuhrman, J.A. (2005) Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. *Limnol Oceanogr* **50:** 620–628.

Schwalbach, M.S., Brown, M., and Fuhrman, J.A. (2005) Impact of light on marine bacterioplankton community structure. *Aquat Microb Ecol* **39:** 235–245.

Sieracki, M.E., Gilg, I.C., Thier, E.C., Poulton, N.J., and Goericke, R. (2006) Distribution of planktonic aerobic anoxygenic photoheterotrophic bacteria in the northwest Atlantic. *Limnol Oceanogr* **51:** 38–46.

Swingley, W.D., Gholba, S., Mastrian, S.D., Matthies, H.J., Hao, J., Ramos, H., *et al.* (2007) The complete genome sequence of *Roseobacter denitrificans* reveals a mixotrophic as opposed to photosynthetic metabolism. *J Bacteriol* (in press). 8

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glockner, F.O. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5:** 163.

Venter, J.C., Remington, K., Heidelberg, J., Halpern, A.L., Rusch, D., Eisen, J.A., *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74.

Waidner, L.A., and Kirchman, D.L. (2005) Aerobic anoxygenic photosynthesis genes and operons in uncultured bacteria in the Delaware River. *Environ Microbiol* **7:** 1896–1908.

Wall, M.K., Mitchenall, L.A., and Maxwell, A. (2004) *Arabidopsis thaliana* DNA gyrase is targeted to chloroplasts and mitochondria. *Proc Natl Acad Sci USA* **101:** 7821–7826.

Walsh, D.A., Bapteste, E., Kamekura, M., and Doolittle, W.F. (2004) Evolution of the RNA polymerase B′ subunit gene (*rpoB′*) in Halobacteriales: a complementary molecular marker to the SSU rRNA gene. *Mol Biol Evol* **21:** 2340–2351.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* (in press). 9

Yurkov, V.V., and Beatty, J.T. (1998) Aerobic anoxygenic phototrophic bacteria. *Microbiol Mol Biol Rev* **62:** 695–724.

Yutin, N., and Béjà, O. (2005) Putative novel photosynthetic reaction center organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from environmental genomics and metagenomics. *Environ Microbiol* **7:** 2027–2033.

Yutin, N., Suzuki, M.T., and Béjà, O. (2005) Novel primers reveal a wider diversity among marine aerobic anoxygenic phototrophs. *Appl Env Microbiol* **71:** 8958–8962.

Supplementary material

The following supplementary material is available for this article online:

Fig. S1. Comparisons between *pufM*, *pufL* and *bchX* GOS data.
A. The reads associated with *pufM* (orange), *pufL* (blue) and *bchX* (green) genes.
B. *pufM* (orange), *pufL* (blue) and *bchX* (green) read equivalents. *X*-axes represent sampling site numbers in the same order as at Figs 4 and 6.

Fig. S2. The use of *gyrA* and *rpoB* genes as alternative bacteria identifiers. Aerobic anoxygenic photosynthetic bacteria abundances normalized by *recA* (orange), *gyrA* (blue) and *rpoB* (green). *X*-axis represents sampling site numbers in the same order as at Figs 4 and 6.

Table S1. GOS sampling site descriptions.

Table S2. Contributions of each GOS station to *pufM*-containing scaffolds revealed in this study.

Doc S1. An example of read equivalent calculation in station 8 (for the *pufM* gene).

This material is available as part of the online article from http://www.blackwell-synergy.com

SNP Best-set Typesetter Ltd.	
Journal Code: EMI	Proofreader: Jason
Article No: 1265	Delivery date: 14 February 2007
Page Extent: 12	Copyeditor: Edmond

AUTHOR QUERY FORM

Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query References	Query	Remark
q1	Au: Please provide the author name(s) for the unpublished data.	
q2	Au: Throughout the article, Figure 6 has been changed to Figure 5, while Figure 5 has been changed to Figure 6, so that they appear in sequence.	
q3	Au: Please clarify what the asterisk means here.	
q4	Au: (Case *et al.* 2007) Please update the volume number and the page range.	
q5	Au: (Koblízek *et al.* 2006) Please update the volume number and the page range.	
q6	Au: (Ludwig et al. 2004) Please confirm the author group is correct.	
q7	Au: (Rusch *et al.* 2007) 'submitted' has been changed to 2007. The 'submitted' paper should not be include in the list unless it has been accepted for publication. Please provide more details if it has been accepted for publication; otherwise, please remove it from the list and cite it in the text only.	
q8	Au: (Swingley *et al.* 2007) Please provide the volume number and the page range if available.	
q9	Au: (Yooseph *et al.* 2007) Please provide the volume number and the page range if available.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	• • • under matter to remain	Ⓥ
Insert in text the matter indicated in the margin	⋏	New matter followed by ⋏ or ⋏ⓧ2
Delete	/ through single character, rule or underline or ├──┤ through all characters to be deleted	δ or δⓧ2
Substitute character or substitute part of one or more word(s)	/ through letter or ├──┤ through characters	new character / or new characters /
Change to italics	— under matter to be changed	⎵
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	~ under matter to be changed	~
Change to bold italic	≂ under matter to be changed	≂
Change to lower case	Encircle matter to be changed	≢
Change italic to upright type	(As above)	⊹
Change bold to non-bold type	(As above)	⩫
Insert ‘superior’ character	/ through character or ⋏ where required	Y or ⅄ under character e.g. Y2 or ⅄2
Insert ‘inferior’ character	(As above)	⋏ over character e.g. ⋏2
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	Y‘ or ⅄‘ and/or Y’ or ⅄’
Insert double quotation marks	(As above)	Y“ or ⅄“ and/or Y” or ⅄”
Insert hyphen	(As above)	├─┤
Start new paragraph	_┌	_┌
No new paragraph	↩	↩
Transpose	⊔⊓	⊔⊓
Close up	linking ⁐ characters	⁐
Insert or substitute space between characters or words	/ through character or ⋏ where required	Y
Reduce space between characters or words	| between characters or words affected	T