The ISME Journal (2010) 4, 784–798

www.nature.com/ismej

ORIGINAL ARTICLE

Genome characteristics of a generalist marine bacterial lineage

Ryan J Newton[1], Laura E Griffin[1], Kathy M Bowles[1], Christof Meile[1], Scott Gifford[1], Carrie E Givens[1], Erinn C Howard[1], Eric King[1], Clinton A Oakley[2], Chris R Reisch[3], Johanna M Rinta-Kanto[1], Shalabh Sharma[1], Shulei Sun[1], Vanessa Varaljay[3], Maria Vila-Costa[1,4], Jason R Westrich[5] and Mary Ann Moran[1]

[1]*Department of Marine Sciences, University of Georgia, Athens, GA, USA;* [2]*Department of Plant Biology, University of Georgia, Athens, GA, USA;* [3]*Department of Microbiology, University of Georgia, Athens, GA, USA;* [4]*Group of Limnology-Department of Continental Ecology, Centre d'Estudis Avançats de Blanes-CSIS, Catalunya, Spain and* [5]*Odum School of Ecology, University of Georgia, Athens, GA, USA*

Members of the marine Roseobacter lineage have been characterized as ecological generalists, suggesting that there will be challenges in assigning well-delineated ecological roles and biogeochemical functions to the taxon. To address this issue, genome sequences of 32 Roseobacter isolates were analyzed for patterns in genome characteristics, gene inventory, and individual gene/pathway distribution using three predictive frameworks: phylogenetic relatedness, lifestyle strategy and environmental origin of the isolate. For the first framework, a phylogeny containing five deeply branching clades was obtained from a concatenation of 70 conserved single-copy genes. Somewhat surprisingly, phylogenetic tree topology was not the best model for organizing genome characteristics or distribution patterns of individual genes/pathways, although it provided some predictive power. The lifestyle framework, established by grouping isolates according to evidence for heterotrophy, photoheterotrophy or autotrophy, explained more of the gene repertoire in this lineage. The environment framework had a weak predictive power for the overall genome content of each strain, but explained the distribution of several individual genes/pathways, including those related to phosphorus acquisition, chemotaxis and aromatic compound degradation. Unassembled sequences in the Global Ocean Sampling metagenomic data independently verified this global-scale geographical signal in some Roseobacter genes. The primary findings emerging from this comparative genome analysis are that members of the lineage cannot be easily collapsed into just a few ecologically differentiated clusters (that is, there are almost as many clusters as isolates); the strongest framework for predicting genome content is trophic strategy, but no single framework gives robust predictions; and previously unknown homologs to genes for H_2 oxidation, proteorhodopsin-based phototrophy, xanthorhodpsin-based phototrophy, and CO_2 fixation by Form IC ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) expand the possible mechanisms for energy and carbon acquisition in this remarkably versatile bacterial lineage.
The ISME Journal (2010) **4,** 784–798; doi:10.1038/ismej.2009.150; published online 14 January 2010
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: genome; marine; phylogeny; Roseobacter

Introduction

Since the discovery of their abundance in marine bacterioplankton communities two decades ago (González and Moran, 1997), members of the marine Roseobacter lineage have emerged as important model organisms for marine microbial ecology. The group spans multiple described genera (at least 45), encompasses a comparatively large sequence variation among 16S rRNA genes (up to 11%) and has a poorly resolved within-taxon phylogeny (Buchan *et al.*, 2005; Wagner-Döbler and Biebl, 2006; Brinkhoff *et al.*, 2008). The recent availability of genome sequences (currently, 5 closed and 27 draft) from cultured members of the Roseobacter lineage provides a detailed inventory of the metabolic and ecological capabilities of each strain (albeit limited by the accuracy of annotation), a basis for comparative analyses among strains, and a means to examine predictive frameworks for the lineage.

Genome sequences of other ocean microbes have been used to explore niches and resource partitioning within taxa. Multiple genome sequences and robust phylogenies for *Prochlorococcus* have revealed that the distribution of ecologically important gene systems (for example, light harvesting

Correspondence: MA Moran, Department of Marine Sciences, University of Georgia, Marine Science Building, Athens, GA 30602-3636, USA.
E-mail: mmoran@uga.edu
Received 5 October 2009; revised 30 November 2009; accepted 2 December 2009; published online 14 January 2010

(Garczarek *et al.*, 2000; Hess *et al.*, 2001; Bibby *et al.*, 2003) and DNA repair mechanisms (Scanlan *et al.*, 2009)) strongly correlate with the phylogenetic structure of this genus (Rocap *et al.*, 2003; Coleman and Chisholm, 2007). In addition, genome content differences among both *Prochlorococcus* and *Synechococcus* strains have been linked to variations in the environments from which the strains were isolated (West and Scanlan, 1999; Johnson *et al.*, 2006; Martiny *et al.*, 2006; Palenik *et al.*, 2006; Dufresne *et al.*, 2008). For members of the Vibrionaceae, small-scale differences in environmental conditions based on microenvironment and season have been shown to drive lineage adaptation (Hunt *et al.*, 2008) and presumably genome content. The phylogenetic- and environment-based frameworks used to interpret data from studies such as these have facilitated the development of predictive community structure models for marine microbes (Follows *et al.*, 2007; Rabouille *et al.*, 2007).

Members of the Roseobacter lineage have been characterized as ecological generalists (Moran *et al.*, 2004, 2007; Polz *et al.*, 2006). Although the first cultured roseobacters were aerobic anoxygenic phototrophs (AAnPs) (Shiba *et al.*, 1979), numerous heterotrophic strains have since been found (Blankenship *et al.*, 1995; Shimada, 1995; Buchan *et al.*, 2005). Cultured roseobacters have a surprisingly flexible suite of mechanisms for energy and carbon acquisition, including carbon monoxide and hydrogen sulfide oxidation (King, 2003; Moran *et al.*, 2004), and anaplerotic CO_2 fixation (Sorokin *et al.*, 2003; Moran *et al.*, 2004; Swingley *et al.*, 2007). Together with the observation that this lineage's genomes are large and variable, a picture of considerable trophic versatility among roseobacters has emerged (Buchan *et al.*, 2005; Wagner-Döbler and Biebl, 2006; Brinkhoff *et al.*, 2008).

The 32 Roseobacter genomes provide an unprecedented opportunity to examine the scope of extant gene systems and to explore various ecological and evolutionary perspectives that might distinguish functionally differentiated clusters within this lineage. In this study, we consider three theoretical ecological/evolutionary frameworks as possible predictors of the gene repertoires of the 32 Roseobacter strains. As a robust Roseobacter phylogeny has yet to emerge from rRNA gene analysis (Buchan *et al.*, 2005; Brinkhoff *et al.*, 2008), we first develop a well-supported phylogeny for the lineage from a concatenation of conserved, single-copy genes and within this phylogenetic structure we examine the evolutionary relationships as possible constraints on genome content and predictors of the genetic capabilities of each strain. Next, we explore lifestyle strategy (heterotroph, photoheterotroph or autotroph) as a possible driver of genome attributes (that is, imposing or releasing bacteria from constraints on genome content). Finally, we ask whether environmental conditions (defined here by the geographical location of isolation) might best explain the observed differences in genetic traits and the retention or acquisition of specific gene systems.

Materials and methods

Genome sequencing, annotation and completeness index

The 32 Roseobacter genomes publicly available as of 15 August 2008 were used in analyses (see Table 1 for genome details). Sequencing and annotation methods for *Ruegeria pomeroyi* DSS-3 (Moran *et al.*, 2004; formerly *Silicibacter*), *Ruegeria* sp. TM1040, *Jannaschia* sp. strain CCS1 (Moran *et al.*, 2007), *Roseobacter denitrificans* OCh 114 (Swingley *et al.*, 2007) and *Dinoroseobacter shibae* DFL12 (Wagner-Döbler *et al.*, 2009) are described elsewhere. The remaining genomes (Table 1) were sequenced and auto-annotated by the J Craig Venter Institute as part of the Moore Foundation Microbial Genome Sequencing Project (see http://moore.jcvi.org/moore/ for details).

In all, 5 of the 32 Roseobacter genomes have been assigned closed genome status, and we used these genomes as the basis for our genome completeness index. The protein sequences of 143 universal single-copy bacterial genes (Santos and Ochman, 2004; Santos, personal communication) were used in a BLASTp query against the five closed Roseobacter genomes; manual gene-calling based on alignment score, *E*-value, and contextual analysis was used to determine the presence or absence of genes in the five closed genomes. Of these 143 genes, 111 were determined to be unambiguously present in all five genomes (see Supplementary Table S1). The protein sequences of these 111 genes were then used in BLASTp analysis against the remaining 27 genomes. The percent presence of these 111 genes in a single genome constituted that genome's completeness index (Table 1).

Phylogenetic tree inference

Out of 111 universal single-copy genes identified in the 32 Roseobacter genomes, only genes that were completely sequenced in all genomes and had no ambiguous start/stop sites were used in phylogenetic analyses. These 70 genes (Supplementary Table S1) were concatenated and aligned with ClustalW in Geneious 4.0 (available from http://www.geneious.com) using *Escherichia coli* K12 substrain MG1655 as the outgroup. The alignment was imported into ARB (Ludwig *et al.*, 2004), where it was heuristically adjusted, and a filter was created to remove all positions containing gaps in the alignment. The resultant alignment of 25 316 positions was used in subsequent phylogenetic reconstruction analyses in ARB (neighbor joining with point accepted mutation substitution matrix and 100 bootstrap runs) and in RAxML (Stamatakis *et al.*, 2008) at CIPRES (http://www.phylo.org; maximum likelihood analysis with 200 bootstrap

Table 1 Roseobacter genome characteristics

Organism	*Clade*[a]	*Isolation source (category)*	*Phototrophy genes*[b]	*Genome size (Mb)*	*rRNA operons (16S/23S)*[c]	*No. of contigs*	*Genome completeness (%)*[d]	*G+C content (%)*
Closed genome								
Dinoroseobacter shibae DFL 12	5	*Prorocentrum lima*, Bay of Tokyo (E)	AAnP	4.35	2/2	6	100	65
Jannaschia sp. CCS1	5	Bodega Head, USA surface water (P)	AAnP	4.40	1/1	2	100	62
Roseobacter denitrificans OCh 114	2	*Enteromorpha linza*, Australia (E)	AAnP	4.13	1/1	5	100	58
Ruegeria pomeroyi DSS-3	1	Coastal Georgia, USA surface water (A)	None	4.60	3/3	2	100	64
Ruegeria sp. TM1040	1	*Pfiesteria piscicda,* Chesapeake Bay (E)	None	4.15	5/5	3	100	60
Draft genome								
Loktanella vestfoldensis SKA53	4	North Atlantic surface water (A)	AAnP	3.06	1-1p/1	14	99	65
Maritimibacter alkaliphilus HTCC2654	None	Sargasso Sea 10 m water (A)	None	4.53	1/1	46	99	64
Pelagibaca bermudensis HTCC2601	3	Sargasso Sea 10 m water (A)	RuBisCO	5.43	4/4	103	98	66
Oceanibulbus indolifex HEL-45	2	North Sea 10 m water (A)	None	4.11	3/3	105	100	59
Oceanicola batsensis HTCC2597	3	Sargasso Sea 10 m water (A)	None	4.44	1/1	23	99	66
Oceanicola granulosus HTCC2516	4	Sargasso Sea 10 m water (A)	None	4.04	4/4	85	100	70
Octadecabacter antarcticus 307	4	McMurdo Sound (H)	Xanthorhodopsin	4.89	2/2	58	100	54
Octadecabacter arcticus 238	4	Offshore Deadhorse, Alaska (H)	Xanthorhodopsin	5.39	2/2	80	96	55
Phaeobacter gallaeciensis 2.10	1	*Ulva lactuca*, Australian waters (E)	None	4.16	4/4	33	100	59
Phaeobacter gallaeciensis BS107	1	*Pecten maximus*, Spanish waters (E)	None	4.23	4/5	24	100	59
Rhodobacterales bacterium HTCC2083	2	Coastal Oregon 10 m water (P)	AAnP	4.02	2/2	20	99	53
Rhodobacterales bacterium HTCC2150	None	Coastal Oregon, surface water (P)	None	3.58	2/2	25	98	49
Rhodobacterales bacterium HTCC2255	None	Coastal Oregon, 10 m water (P)	Proteorhodopsin	4.81	0-1p/0	70	96	38
Rhodobacterales bacterium Y4I	1	Coastal Georgia, USA water (A)	None	4.33	4/4	63	99	64
Roseobacter litoralis Och149	2	Seaweed (E)	AAnP	4.68	1/1	27	99	57
Roseobacter sp. AzwK-3b	3	Estuary Monterey Bay water (P)	AAnP	4.18	2/2	31	100	61
Roseobacter sp. CCS2	4	Bodega Head, USA surface water (P)	AAnP	3.50	1/1	11	99	55
Roseobacter sp. GAI101	2	Coastal Georgia, USA water (A)	None	4.25	4/4	67	99	58
Roseobacter sp. MED193	1	NW Mediterranean 1 m water (A)	None	4.65	1-1p/1-4p	19	100	57
Roseobacter sp. SK209-2-6	1	Arabian Sea O_2-min., 267 m water (I)	None	4.56	5/5	29	100	57
Roseovarius nubinhibens ISM	3	Caribbean Sea surface water (A)	None	3.67	2/2	10	100	63
Roseovarius sp. 217	3	Coastal England surface water (A)	AAnP	4.76	1-1p/1-1p	37	100	60
Roseovarius sp. TM1035	3	*Pfiesteria piscicda,* Chesapeake Bay (E)	AAnP	4.21	3/3	15	100	60
Ruegeria sp. R11	1	*Delisea pulchra* pathogen, Australia (E)	None	3.82	4/4	17	98	60
Sagittula stellata E-37	3	Coastal Georgia, USA water (A)	None	5.26	2/2	39	98	65
Sulfitobacter NAS-14.1	2	Coastal Georgia, USA surface water (A)	None	4.00	4/4	27	100	60
Sulfitobacter sp. EE-36	2	North Atlantic surface water (A)	None	3.54	4/4	15	100	60

Abbreviations: A, Atlantic Ocean; E, Eukaryote associated; H, polar ocean (high latitude; Arctic or Antarctic); I, Indian Ocean; P, Pacific Ocean.
[a]Defined by the phylogenetic tree shown in Figure 1.
[b]Genomes were defined as aerobic anoxygenic phototroph (AAnP) if the genome contained the *puf* operon and bacteriochlorophyll *a*; as none if the genome contained no light-harvesting genes; as xantho/proteorhodopsin if the genome contained the specific rhodopsin gene; and as RuBisCO if the genome contained a RuBisCO homolog and homologs to the Calvin–Benson–Bassham cycle.
[c]For example, 1-1p indicates that this organism has one fully sequenced rRNA gene and one partially sequenced rRNA gene.
[d]Genome completeness was defined by examining 111 universal genes found in all closed Roseobacter genomes (see Materials and methods).

runs and Jones–Taylor–Thornton (JTT) substitution model). The best-fit maximum likelihood tree is reported along with bootstrap values from each phylogenetic inference method.

Identification of orthologs and ecologically relevant genes

Orthologs among the 32 genomes were identified by sequential two-way reciprocal best-hit (RBH) analysis, beginning with the *R. pomeroyi* and Rhodobacterales HTCC2255 genome comparison and continuing by adding each of the remaining 30 genomes one at a time. The RBH Basic Local Alignment Search Tool (BLAST) thresholds were set at *E*-value $<10^{-5}$ and amino acid identity $>30\%$. The RBH results were subsequently compiled into a single matrix containing the distribution of all shared genes and used for the genome content comparisons described below (See Supplementary Table S2 for matrix). This relaxed ortholog definition was used because an all-way RBH requirement was unworkable for the large number of genomes, each containing gene families represented by multiple members. We tested whether the order of the sequential best-hit analysis resulted in substantial changes in the ortholog matrix or the outcome of the analyses (that is, by using a different order of adding genomes in the pair-wise RBH), and found it did not.

In addition to whole-genome ortholog identification, a select group of ecologically relevant genes/gene pathways was also identified using representative protein sequences of the target genes from a Roseobacter for which the gene functions had been experimentally verified. If no Roseobacter met this criterion, then a protein sequence was obtained from the closest Roseobacter relative containing the desired experimentally verified gene. All query protein sequences were used in BLASTp analysis against the Roseobacter genome database (http://www.roseobase.org). BLAST *E*-values, gene neighborhoods and clusters of orthologous group assignments (Tatusov *et al*., 2003) were manually examined and used to determine the presence or absence of these genes and pathways in each of the 32 genomes.

Classification schemes

The 32 isolate genomes were sorted into groups within each of the three frameworks. First, five deeply branching nodes in our phylogenetic inference best-fit tree were chosen to distinguish isolate groups based on shared ancestry and were designated Clades 1–5. Next, we categorized isolates into lifestyles based on their trophic status: heterotrophic, photoheterotrophic (that is, heterotrophic but likely subsidized by aerobic anoxygenic phototrophy or rhodopsin-based phototrophy) or autotrophic. Organisms were considered AAnPs based on the presence of the *puf* operon and genes for the synthesis of bacteriochlorophyll *a*; they were considered rhodopsin-supplemented photoheterotrophs based on the presence of gene orthologs for proteorhodopsin or xanthorhodopsin; and they were considered autotrophs based on the presence of RuBisCO and the Calvin–Benson–Bassham pathway. Although these designations were made from draft genome sequences for many strains, the high genome completeness index suggests they are largely correct. Finally, isolates were classified into one of five broad environmental categories based on the source of isolation: Pacific Ocean, Atlantic Ocean, Indian Ocean, polar oceans or eukaryote-associated (Table 1).

Identification of genes in the Global Ocean Sampling

Global Ocean Sampling (GOS) sample sites, sampling procedures and sequencing methods are described elsewhere (Rusch *et al*., 2007; Yooseph *et al*., 2007). A subset of Roseobacter protein sequences representing each of the major biogeochemical pathways and processes that we examined (Supplementary Table S3) was used in a BLASTp query against the unassembled GOS data set at the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) (Seshadri *et al*., 2007). Gene hits from the GOS samples were retained for further analysis as potential orthologs at *E*-value cutoffs ranging between 10^{-80} and 10^{-20}, depending on the gene. Paired reads were then removed before the resultant matches were used in BLAST analysis against the All Prokaryotic Proteins (P) database. Only gene matches that had a best hit to a gene in a Roseobacter genome were retained for further analysis. Finally, protein sequences from the Roseobacter-like GOS matches underwent a BLASTp query at GenBank, and were eliminated if their top alignment scores were to proteins with a different annotated function than the original query protein.

To compare gene counts between oceans, the Roseobacter-like metagenomic sequences obtained from the GOS data underwent several normalizations. Counts for functional genes retrieved at each sample location were size-normalized to the length of the *recA* gene from *E. coli* K12 substrain MG1655 to account for effects of size on the probability of sampling (Howard *et al*., 2008). The number of Roseobacter genome equivalents for each sample location was then calculated by averaging size-normalized Roseobacter-like gene counts of the universal single-copy genes *recA* and *rpoB*. To estimate per-cell frequency for each examined Roseobacter gene (listed in Supplementary Table S3), the sample gene counts were summed by ocean basin (Atlantic, Pacific and Indian) and divided by the number of Roseobacter genome equivalents for that basin. Only coastal and open ocean GOS sample sites were considered, with estuaries, embayments, lagoon reefs, fringing reefs, freshwater, mangroves, coral reefs, hypersaline lagoons, warm seeps and

harbors excluded from the analyses (Supplementary Table S4).

The clade distribution among ocean basins was determined using the *recA* gene sequence. After retrieving Roseobacter-like RecA sequences by BLASTp analysis against the GOS database at CAMERA (as described above), each individual protein sequence was used as a query sequence in a subsequent BLASTp analysis against all 32 genomes at Roseobase (http://www.roseobase.org). The Roseobacter-like RecA sequences from GOS were then assigned to a clade according to their best match among the 32 genomes, and the occurrences of each clade were summed across samples in each ocean basin.

Statistical analyses

Patterns of ortholog distribution among the 32 genomes were evaluated using the Bray–Curtis Index of Similarity (Legendre and Legendre, 1998). The 32 genomes contained a total of 31 874 orthologs. Similarities between genomes include all orthologs in this matrix, so that both the shared presence and shared absence of a gene are taken into account in the similarity calculation. This similarity matrix was used to create a hierarchical clustering dendrogram based on complete linkage grouping (that is, furthest neighbor analysis). An analysis of similarity (ANOSIM) was used to test for significant differences among *a priori* assigned genome groups based on phylogenetic clade, trophic strategy or geographical isolation location. The multivariate analyses were performed using the statistical package PRIMER 5 for Windows v. 5.2.7.

The average nucleotide identity was obtained for two Roseobacter strain comparisons, *Phaeobacter gallaeciensis* 2.10 with *P. gallaeciensis* BS107 and *D. shibae* DFL12 with *P. gallaeciensis* 2.10 according to the method described by Goris *et al.* (2007). These comparisons were chosen to bracket the amount of sequence heterogeneity observed and to provide context for our ortholog similarity comparisons.

Significance of gene distributions between any two assigned groups (for example, between two clades, between two trophic strategies or between two ocean basins) was assessed with a binomial distribution *d*-score test (Markowitz *et al.*, 2008).

Results and discussion

Because many of the 32 Roseobacter genomes are in draft status, we developed a completeness index based on the presence or absence of 111 universal single-copy genes. The lowest genome completeness index obtained was 96% (for *Octadecabacter arcticus* 238 and Rhodobacterales bacterium HTCC2255; 107 out of 111 presumed universal genes were represented), and 18 of 32 genomes had a completeness index of 100% (Table 1). We therefore considered all draft genomes to be good representations of these organisms' gene content.

Phylogenetic inference and clade distribution

Previous phylogenetic reconstructions of the Roseobacter lineage using 16S rRNA gene relationships have led to the identification of subgroups within the lineage (Buchan *et al.*, 2005; Brinkhoff *et al.*, 2008). However, many of the nodes, especially those distinguishing deep branching points in these phylogenies, do not have statistical support, and therefore do not provide clear phylogenetic relationships for the members of this lineage. We took advantage of the genome sequence data to construct an alignment from the concatenation of 70 conserved single-copy genes; this alignment was subsequently used in phylogenetic tree inference (see Supplementary Table S1 for gene list and Supplementary Figure S1 for concatenated gene, 16S rRNA gene and 23S rRNA gene tree comparisons). The resultant tree topology suggested there are five deeply branching clades within the Roseobacter lineage (Figure 1). Three of the presumed roseobacters, *Maritimibacter alkaliphilus* HTCC2654, Rhodobacterales HTCC2150 and Rhodobacterales HTCC2255, fell outside these clades. Most members within a single genus clustered together on the tree, although the placement of two members of the genus *Oceanicola* into different clades suggests that a taxonomic reclassification may be needed for some isolates.

Buchan *et al.* (2005) identified 13 major sequence clusters based on 16S rRNA gene sequences within the Roseobacter lineage. Twelve of the 16S rRNA-based clusters can be mapped onto our 70-gene phylogeny (data not shown). Clade 1 contains the 16S rRNA gene sequence clusters RGALL, RATL and TM1040. Clade 2 contains sequence clusters ANT9093, OBULB, SPON and AS-21. Clade 3 contains sequence cluster CHAB-I-5. Clade 4 contains sequence clusters AS-26, DG1128, DC5-80-3 (RCA cluster) and OCT. Clade 5 contains no previously identified sequence clusters, and cluster NAC11-7 is not covered by any of the clades in our study.

Two of the most abundant Roseobacter 16S rRNA gene sequence clusters recovered from marine habitats do not have closely associated sequenced genomes (Buchan *et al.*, 2005), and thus are not included in the 70-gene phylogenetic tree. The first, the DC5-80-3 or RCA cluster, has often been observed as the most abundant Roseobacter group in polar and temperate oceans (Brinkhoff *et al.*, 2008). 16S rRNA genes from RCA distantly group with those from genomes in Clade 4 (Figure 1), a clade that harbors all the sequenced polar Roseobacter isolates thus far. A second abundant marine sequence cluster, NAC11-7, is frequently the dominant Roseobacter taxon found during phytoplankton blooms (Buchan *et al.*, 2005; West *et al.*, 2008). 16S

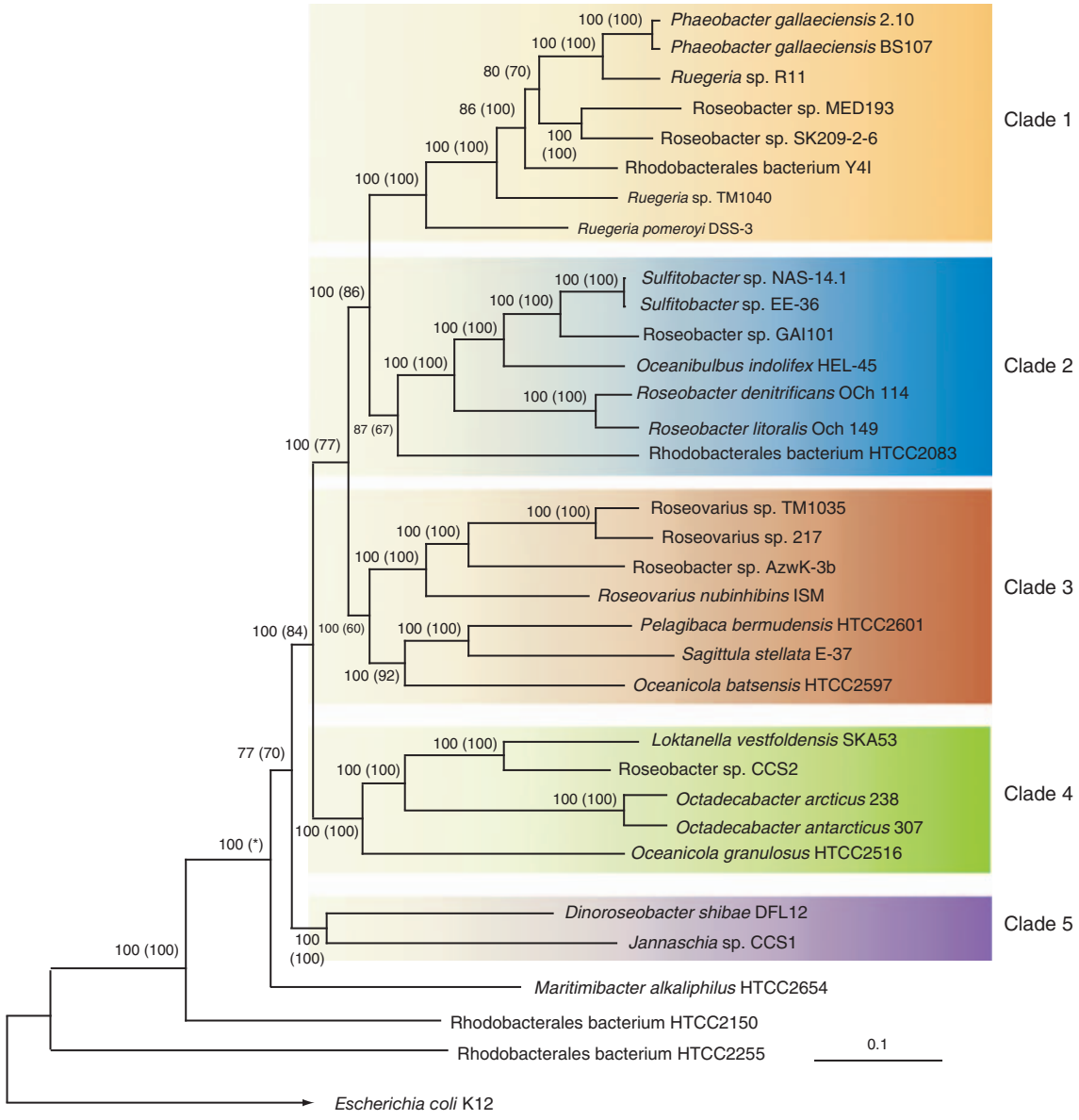

Figure 1 A consensus maximum likelihood tree of the 32 sequenced Roseobacter genomes. The alignment for tree inference was created from a concatenation of 70 universal single-copy genes contained in each of the Roseobacter genomes and in *E. coli* K12, which was used as an outgroup. Bootstrap values of >50% for the maximum likelihood best-fit tree (200 iterations) and neighbor-joining tree (100 iterations) are listed at each node. The neighbor-joining bootstrap values are listed in parentheses. (*) demarcates nodes where the neighbor-joining tree did not agree with the maximum likelihood tree. Designated Clades 1–5 are listed to the right of the tree. The scale bar represents 10% sequence divergence.

rRNA genes from the NAC11-7 group did not cluster with those of any clade, and were most related to that of Rhodobacterales HTCC2255 (data not shown), which also fell outside the five clades established by the 70-gene phylogeny (Figure 1).

Roseobacter-like *recA* genes, a robust marker for bacterial phylogeny (Eisen, 1995), were obtained from the GOS data set by BLASTp analysis (see Materials and methods) to ascertain which isolate genomes are most representative of wild roseobacters in surface ocean water. When the set of Roseobacter-like RecA GOS sequences was used in a best-match BLASTp query against the 32 genomes, hits to all five clades were found throughout the major ocean habitats surveyed (Figure 2). In general, the distribution of clades is not remarkably different between the Atlantic and Pacific, or Indian oceans (Figure 2). A large percentage of the Roseobacter RecA sequences from the GOS appear most closely related to one of the three singleton genomes (that is, not belonging to one of the five defined clades). This finding, along with the lack of genomic data for the RCA and NAC11-7 sequence clusters, suggests that representation of oceanic Roseobacter genomes could be improved with additional genome sequences.

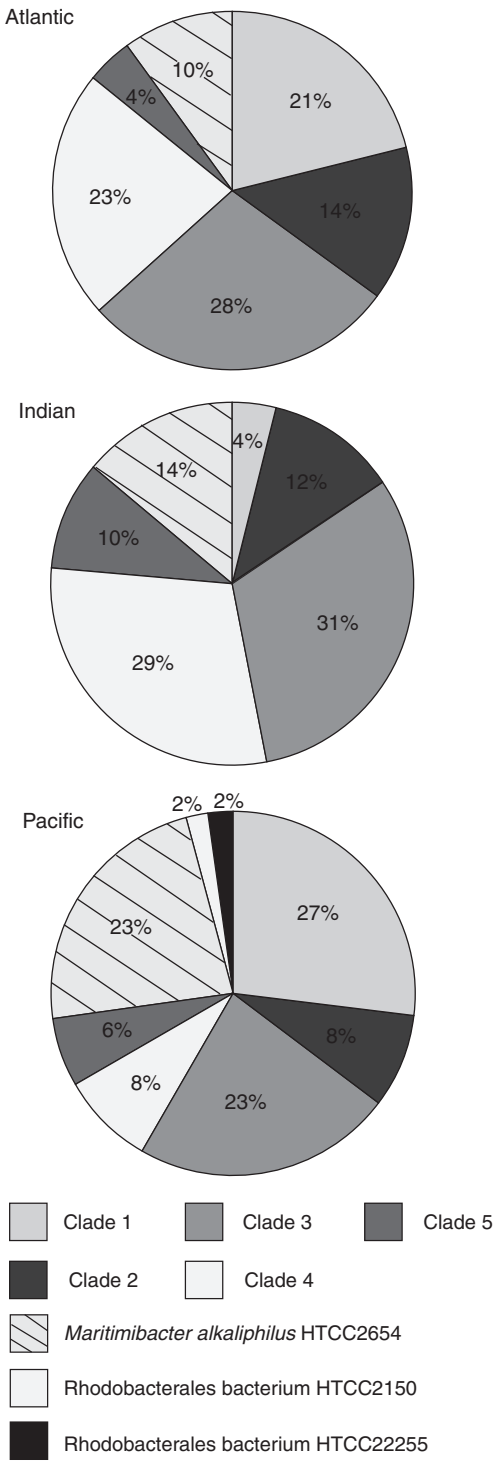

Figure 2 Clade distribution among ocean basins calculated from BLASTp best match of Roseobacter *recA* genes from the Global Ocean Sampling (GOS). Atlantic $n=71$, Pacific $n=48$ and Indian $n=51$.

Genome content related to phylogeny

We examined genome characteristics (for example, G+C content, rRNA copy number, genome size; Table 1) and gene content of the 32 roseobacters within the context of the five-clade phylogenetic framework. Whole genomic content comparisons (based on distribution patterns of 31 874 orthologs; see Materials and methods; Supplementary Table S2) indicate weak but significant genome clustering by clade (ANOSIM $R=0.410$, $P\leqslant 0.001$, three roseobacters not assigned to clades were excluded from the statistical test), with the within-clade similarity in gene repertoire for Clades 1, 2 and 3 driving this pattern (Figure 3). Generally, neither the examined genome characteristics nor the examined gene distributions segregate strongly based on phylogenetic relatedness (Table 1 and Figure 4). Some exceptions include a greater mean rRNA operon copy number for Clade 1 than for other clades (*t*-test, $P\leqslant 0.01$); a strictly heterotrophic composition of Clade 1; a genetic potential for biotin synthesis in Clade 1 (vitamin synthesis in bacteria has been identified as important in bacterial–phytoplankton relationships; Croft *et al.*, 2005; Wagner-Döbler *et al.*, 2009); a lack of Lux-type quorum sensing genes in Clades 4 and 5; a genetic potential for H_2 oxidation unique to Clade 3; the absence of sulfur oxidation genes in Clade 4 genomes; and absence of the *ppk1* gene for polyphosphate biosynthesis in Clade 1 (whereas all isolates outside Clade 1 have this gene).

The lack of a strong segregation by phylogenetic assignment for genome content (Figure 3) or ecologically relevant gene systems (Figure 4) suggests the importance of gene acquisition by horizontal transfer originating either within or outside the lineage. Other evolutionary processes known to shape genome content (selective gene loss, gene duplication, gene genesis; Snel *et al.*, 2002) are no doubt important in this lineage, but are mechanisms less likely to produce the observed patchy distribution of ecologically relevant genes in the Roseobacter isolates relative to their phylogenetic reconstruction. Although the rates of gene transfer within the Roseobacter lineage is not known, the occurrence in 30 of 32 genomes of gene transfer agent operons (Figure 4), an unusual system for moving chromosomal fragments to close relatives (Biers *et al.*, 2008; Zhao *et al.*, 2009), suggests a mechanism for shaping Roseobacter gene content through frequent within-lineage gene transfers.

Between-genome similarities were generally higher for Roseobacter isolates in the same genus (for example, *P. gallaeciensis* BS107 and *P. gallaeciensis* 2.10; *R. denitrificans* Och114 and *Roseobacter litoralis* Och149; Figure 3) than for roseobacters belonging to different genera. Nonetheless, blurred gene content boundaries among deeply branching clades would impose a requirement of dozens of taxonomically shallow groups (for example, species level) to accurately represent Roseobacter contributions to ecosystem functions, thus making the phylogenetic framework a cumbersome approach for defining ecological subgroups.

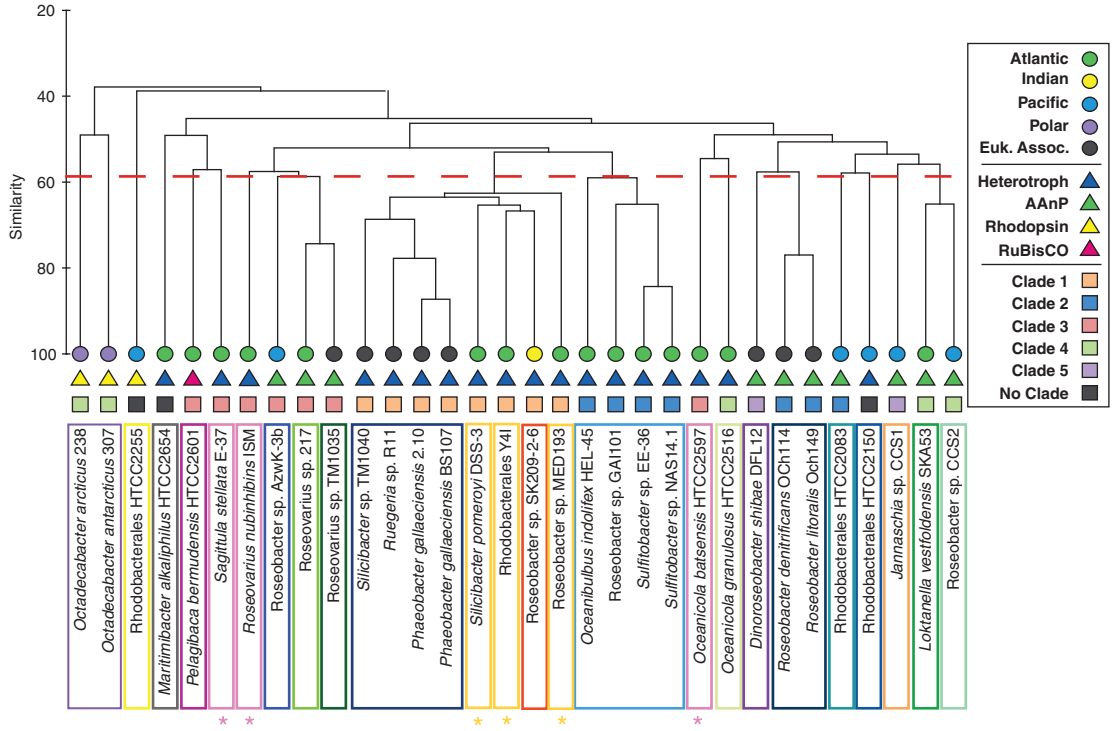

Figure 3 Complete linkage (that is, furthest neighbor) cluster analysis illustrating the gene content similarities among the genomes. Bray–Curtis similarities between all genome pairs were calculated from a matrix containing all 31 874 genes identified in the 32 genomes. In this manner, the similarity calculation was based on both the shared presence and shared absence of genes. For context, the *P. gallaeciensis* 2.10 to *P. gallaeciensis* BS107 comparison is 87.4% similar in this analysis compared with an average nucleotide identity (ANI) (Goris *et al.*, 2007) of 97.0%. The *D. shibae* DFL12 to *P. gallaeciensis* 2.10 comparison is 46.7% similar here compared to an ANI of 70.4%. The three framework groups of each isolate are illustrated by the shape and color pattern depicted at the tips of the cluster diagram. The phylogenetic clade framework is represented by squares; the lifestyle framework is represented by triangles; and the environment framework is represented by circles. Unique combinations of these three frameworks are illustrated with colored boxes around the names of isolates. Breaks in the three framework groupings are noted by an asterisk next to the strain name. Nodes below the dashed red line indicate groups with $\geqslant$58% similarity.

Genome content related to trophic strategy

Owing to the significant versatility in mechanisms for obtaining carbon and energy previously observed for this group (Buchan *et al.*, 2005; Moran *et al.*, 2007), we hypothesized that an organism's trophic strategy could impose or remove constraints on genome content. For example, the ability to use sunlight for energy generation, which is widely distributed within the Roseobacter lineage, might mitigate an organism's energy limitations in the oligotrophic marine environment, while imposing requirements for metals and cofactors specific to phototrophy. Similarly, the ability to fix inorganic carbon might reduce an organism's requirements for substrate transporters. If such interplay between trophic strategy and functional gene repertoire exists, then significant and predictable differences in genome content should be evident between lifestyle categories.

Thirteen of the 32 roseobacters have genes for photoheterotrophy (10 AAnPs, 3 rhodopsin-containing), whereas one has RuBisCO. The remaining 18 are considered heterotrophs here (although some may obtain energy from inorganic compounds such as CO and H_2S; Moran *et al.*, 2007) (Table 1). Genome ortholog comparisons suggest moderate and significant differences in genome content among these groups (ANOSIM $R = 0.545$, $P \leqslant 0.001$). The strength of these differences does not stem solely from the very unique rhodopsin-containing genomes (Figure 3). The differences also are not solely due to the presence of light-harvesting-related genes shared by the AAnP genomes or rhodopsin-containing genomes, as removal of the rhodopsin genes and 29 genes specific for AAnP light harvesting resulted in a similar level of clustering by trophic strategy (ANOSIM $R = 0.522$, $P \leqslant 0.001$). The lifestyle framework accurately predicts the gene repertoire groupings at similarity levels $\geqslant$58% (Figure 3; red dashed line), which represents the gene content relationships for 19 of the 32 genomes and is the best predictor of the three frameworks analyzed.

The majority of non-light-harvesting gene or pathway-related differences among strains can be traced to hypothetical proteins unique to the AAnPs or heterotrophs, as well as to a number of genes encoding transcriptional regulators and amino acid uptake and synthesis systems (Figure 5). Although trophic strategy was a good predictor of an isolate's

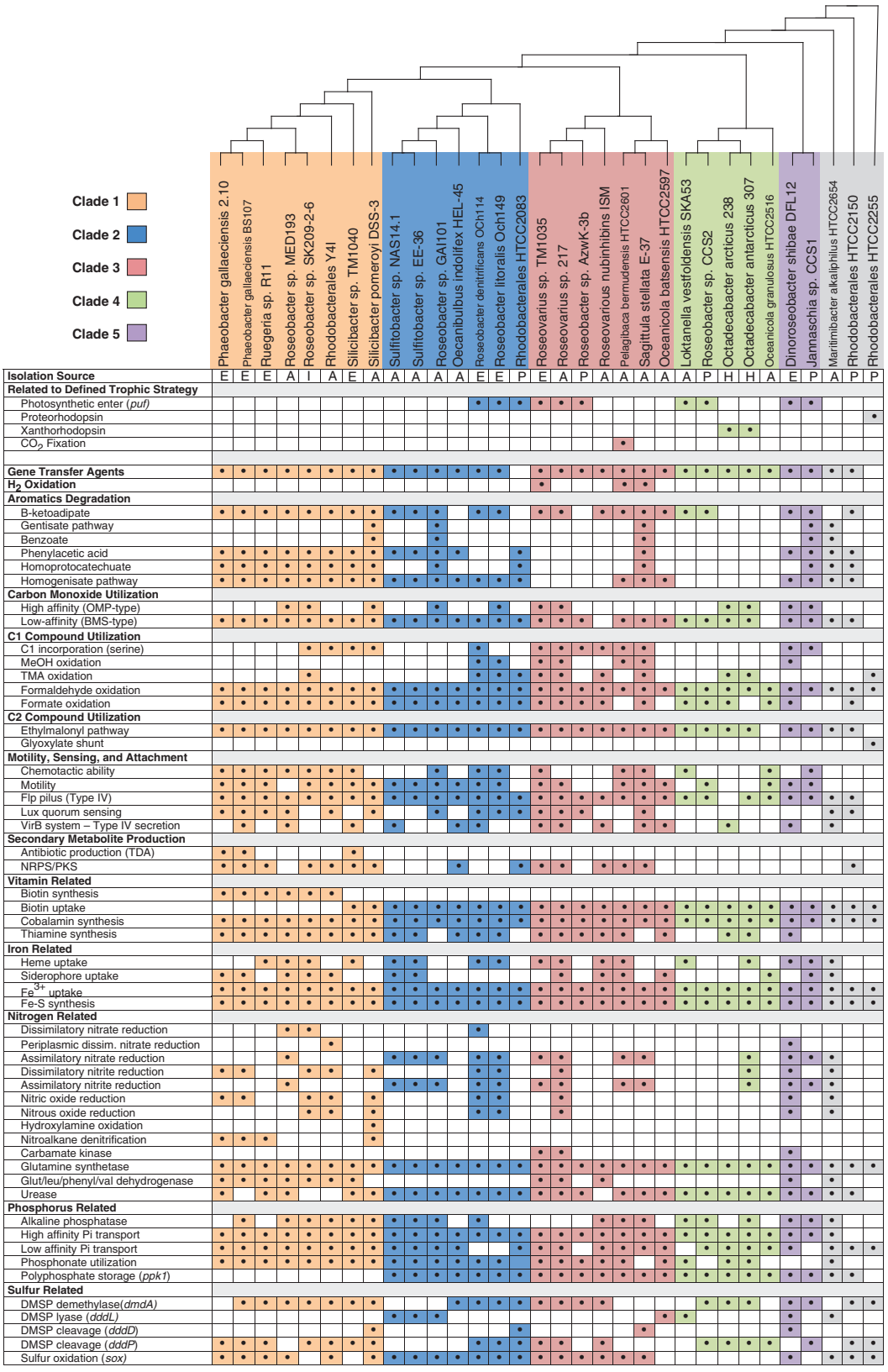

Figure 4 A matrix depicting the presence of select genes or gene pathways in the 32 Roseobacter genomes arranged and color-coded by clade. A colored box containing a dot indicates the presence of the gene/pathway. An ultrametric tree has been placed above the gene matrix for reference. Isolation source indicates the region where the Roseobacter strain was isolated and is coded as: A = Atlantic Ocean, E = Eukaryote Associated, H = polar oceans (high latitude), I = Indian Ocean and P = Pacific Ocean. Gene/pathway abbreviations are as follows: NRPS/PKS, non-ribosomal peptide synthetase/polyketide synthase; Glut/leu/phenyl/val dehydrogenase, glutamate/luecine/phenylalanine/valine dehydrogenase.

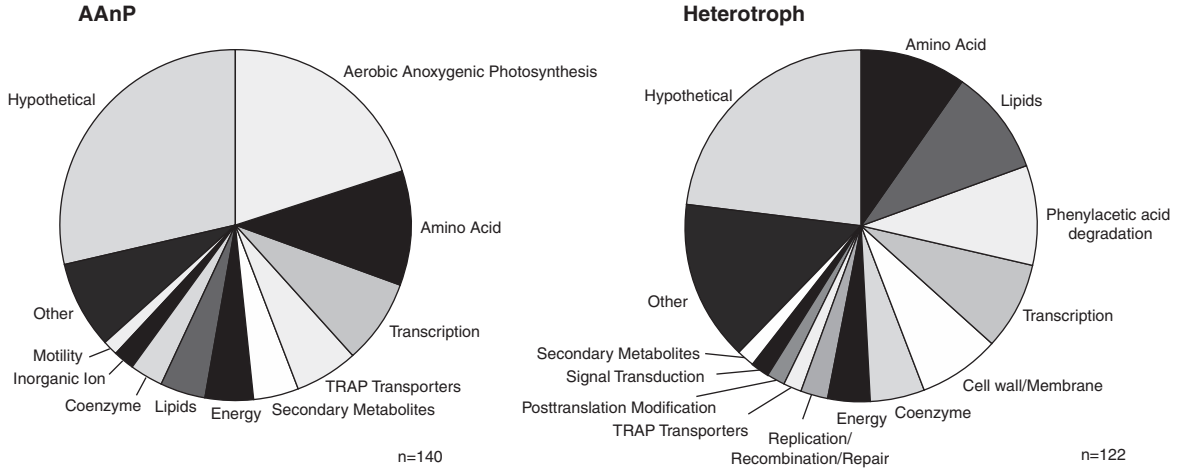

Figure 5 Relative abundance of ortholog groups that are overrepresented in the isolate genomes from a particular lifestyle strategy. An ortholog was considered overrepresented when it was $\geqslant$50% more prevalent in the genomes from one lifestyle strategy than the other. The overrepresented orthologs were grouped into functional categories whose relative percent abundance is depicted. The rhodopsin-containing and RuBisCO-containing lifestyle groupings were not considered because of the low number of genomes in these categories.

gene repertoire, only a few of the ecologically relevant gene systems we examined were strictly differentiated according to this framework. The five C1 utilization pathways identified in Roseobacter genomes (serine cycle, methanol oxidation, trimethylamine oxidation, formaldehyde oxidation and formate oxidation) had a 70% occurrence rate in AAnP genomes (that is, 35 out of the 50 possible occurrences if all 10 AAnP genomes had all five pathways), but only a 42% occurrence rate in the 19 heterotrophs (only 40 out of 95 possible occurrences; *d*-score, $P \leqslant 0.01$). The heterotrophs tended to have more genes for six identified aromatic degradation pathways (β-ketoadipate, gentisate, benzoate, phenylacetic acid, homoprotocatechuate and homogenisate) with a 60% occurrence rate (68 out of 114 possible occurrences) compared with 33% in AAnPs (20 of 60 possible occurrences; *d*-score, $P \leqslant 0.01$). Compared with the genomes of the other groups, the rhodopsin-containing genomes shared few orthologs that distinguished them as a coherent group (data not shown).

As noted previously (Buchan *et al.*, 2005; Moran *et al.*, 2007) and strongly reinforced in this analysis, Roseobacter genomes exhibit a remarkably versatile suite of mechanisms for energy and carbon acquisition. Along with the presence of genes for oxidizing carbon monoxide and hydrogen sulfide (King, 2003; Moran *et al.*, 2004), we found evidence for energy generation by H_2 oxidation in *Roseovarius* sp. TM1035, *Pelagibaca bermudensis* HTCC2601, and in *Sagittula stellata* E-37, proteorhodopsin- (Rhodobacterales sp. HTCC2255) and xanthorhodopsin-based (*O. antarcticus* 307 and *O. arcticus* 238) phototrophy, and CO_2 fixation based on the presence of a Form IC RuBisCO and homologs to Calvin–Benson–Bassham cycle genes in *P. bermudensis* HTCC2601 (Figure 4). This emerging picture of high trophic versatility among cultured roseobacters (Buchan *et al.*, 2005; Wagner-Döbler and Biebl, 2006; Brinkhoff *et al.*, 2008) is in accord with recent shifts away from a perception of marine bacterioplankton communities consisting largely of canonical photosynthetic and heterotrophic cells (Karl, 2002).

The rhodopsin-containing genomes

The three rhodopsin-containing genomes harbored the most unique genome content of any of the isolates (Figure 3). The proteorhodopsin-containing Rhodobacterales bacterium HTCC2255 gene content was unique because it consisted of only 2197 genes, far fewer than any the other genome. The two xanthorhodpsin-containing isolates, *O. arcticus* and *O. antarcticus* (which are also the only two polar ocean isolates), are clearly part of the Roseobacter lineage (Figure 1) but possess the greatest number of unique genes among all isolates (2230 genes for *O. arcticus* and 1822 genes for *O. antarcticus*; 32 isolate mean = 617 and s.d. = 437). The majority of these unique genes were annotated as phage or transposase genes with 65 and 52 phage gene annotations and 935 and 574 transposase gene annotations for *O. arcticus* 238 and *O. antarcticus* 307, respectively; these numbers are extremely high compared with those in the other Roseobacter genomes (phage mean = 21 and s.d. = 10; transposase mean = 52 and s.d. = 38).

Genome content related to ocean environment

Environmental properties are potential drivers of marine bacterial genome evolution by selecting for niche-specific genetic capabilities. For example, the marine cyanobacteria *Prochlorococcus* and *Synechococcus* exhibit gene content patterns that correlate well with the geographical locations from

which they were isolated (Rocap *et al.*, 2003; Palenik *et al.*, 2006; Zwirglmaier *et al.*, 2008; Martiny *et al.*, 2009). For the third framework, we divided the Roseobacter strains into five broad environmental categories that resulted in the following distribution: 15 isolates from the Atlantic Ocean, 1 isolate from the Indian Ocean, 6 isolates from the Pacific Ocean, 2 isolates from the polar oceans and 8 isolates that were cultured in association with eukaryotic organisms (Table 1). We hypothesized that characteristics distinguishing major environments (for example, access to nutrients, differences in temperature) would exert a detectable influence on gene patterns in the Roseobacter genomes.

There was a significant but very weak relationship between genome content and environmental origin of the isolate (ANOSIM $R=0.296$, P-value $\leqslant 0.002$), although if only the Atlantic Ocean and Pacific Ocean isolates were compared, this relationship was stronger (ANOSIM $R=0.398$, P-value $\leqslant 0.002$). Examination of ortholog patterns suggests that the relationship is based on the genomic distribution of motility genes, with 33% occurrence in Pacific Ocean isolates compared with 73% in Atlantic Ocean isolates; chemotaxis genes, with 17% occurrence in Pacific Ocean isolates compared with 47% in Atlantic Ocean isolates; denitrification systems, with 5% in Pacific Ocean isolates compared with 29% in Atlantic Ocean isolates; phosphorus uptake systems known to function at low phosphate concentrations (alkaline phosphatases, high-affinity phosphate uptake and phosphonate uptake), with 50% occurrence (12 out of 24 possible) in Pacific Ocean isolates compared with 85% (51 out of 60 possible) for Atlantic Ocean isolates; and aromatic carbon degradation pathways (mixed ocean basin patterns depending on the specific pathway) (Figure 4, all comparisons d-score P-value $\leqslant 0.01$). There were also a number of unique carbon and ion transporters, amino-acid metabolism genes, and transcription regulators restricted to each ocean basin, and a large suite of unique genes shared by the two polar isolates (data not shown). We tested whether the higher frequency of coastal strains among the Pacific isolates compared with the Atlantic (Table 1) was the basis for the apparent ocean basin pattern, but found it not to be the case whether comparing whole-genome ortholog patterns (ANOSIM $R=0.031$, $P=0.23$ for coastal vs open ocean isolate comparison) or individual gene systems (Supplementary Figure S2).

Geographical patterns in the GOS data set

Despite the many factors that might obscure large-scale environmental imprints (including varied isolation methods, isolation dates spanning several decades and sparse spatial coverage), the ocean basin of isolation seemingly had predictive power for the distribution of select genes/pathways among the Roseobacter genomes. To determine whether this apparent grouping of genome content by geographical origin applies broadly to populations of roseobacters in the world oceans, we probed the GOS data set for similar environmental patterns. As the other two frameworks (phylogeny and lifestyle strategy) require assembled genomes, it is not possible to test for these among the GOS Roseobacter populations.

Homologs to genes listed in Supplementary Table S3 were identified in the GOS peptide sequence database (which currently does not include polar ocean metagenomic data). They were designated as Roseobacter homologs if they had greatest similarity to a gene in a Roseobacter genome in subsequent BLASTp query analysis against all available bacterial genome sequences (the CAMERA 'All Prokaryotic Proteins (P) database'). Many of the same patterns in gene distribution found for cultured roseobacters were evident in the metagenomic analysis (Figure 6a). Most notable was that all phosphorus acquisition systems known to function at low phosphate concentrations (alkaline phosphatases, high-affinity phosphate uptake and phosphonate uptake) were much more abundant in wild roseobacters from the Atlantic Ocean, where the mean phosphate concentration is lower, than for either the Indian or Pacific Ocean (mean phosphate concentration is 0.06 μM for the Atlantic vs 0.15 μM for the Indian vs 0.53 μM for the Pacific; see Martiny *et al.*, 2009 for details). The phosphate uptake system (*pitA*), which operates at high phosphate concentrations, had the opposite pattern, being more abundant in the Indian and Pacific Oceans (Figure 6a). Recently, other studies have noted similar trends for phosphorus gene distribution in the *Prochlorococcus* and SAR11 lineages (Rusch *et al.*, 2007; Martiny *et al.*, 2009), indicating that phosphorus concentration may impart a strong selective force on marine bacterial genomes. Of particular note were Roseobacter genes encoding for phosphonate uptake and assimilation, which exhibited a very large bias in distribution toward the Atlantic Ocean (Figure 6a).

Most representative genes we examined were more prevalent in the isolate genomes than in our per-genome-equivalent calculations for the GOS samples (Figure 6b), an observation that cannot be attributed to sampling disparities as 158 Roseobacter genome equivalents were sampled in the GOS (see Materials and methods). Compared with the roseobacters represented in culture, natural Roseobacter populations in the ocean are more likely to have genes for processing DMSP and utilization of C1 carbon compounds, but less likely to have genes involved in motility, adhesion, quorum sensing, gene transfer and iron uptake (Figure 6b). The higher prevalence of selected genes in the isolate genomes compared with GOS samples may indicate that there are fewer genes per genome in wild cells, could be indicative of the differences in sampling locations between the GOS samples and the isolates, or might reflect a bias during our analysis in selecting genes previously noted in cultured Roseobacter genomes.

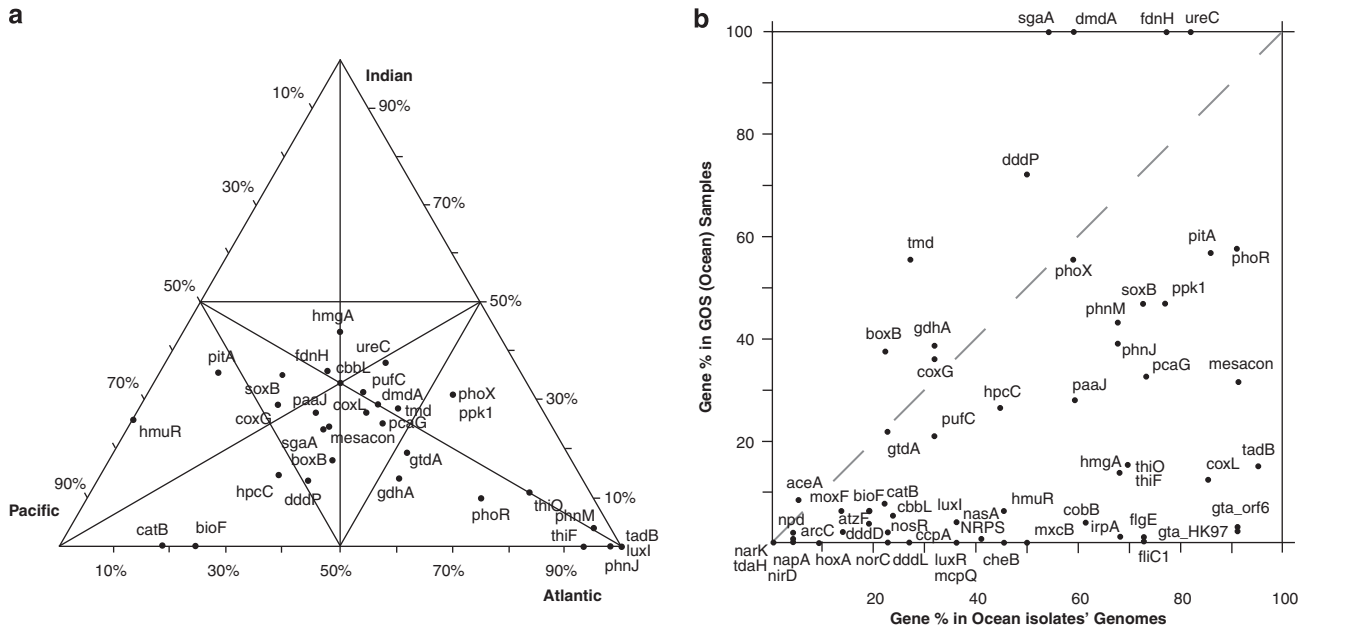

Figure 6 (**a**) Ocean basin (Atlantic, Indian and Pacific) three-way comparison of Roseobacter genes/gene pathways (see Figure 4). The position of each dot indicates the relative abundance of a gene in the Global Ocean Sampling (GOS) data set, based on 61, 52 and 45 Roseobacter genome equivalents in the Atlantic, Indian and Pacific Ocean data sets, respectively. Only genes present at a frequency $\geqslant$10% of genome equivalents in any one ocean basin are depicted. Triangle vertices indicate 100% relative abundance of that particular gene in the representative ocean. Three lines creating an inverted triangle have been drawn to aid in visual interpretation and indicate a relative abundance = 50% in a single ocean basin. (**b**) Gene occurrence percentage of Roseobacter genome equivalents in the GOS data set ($n = 158$) vs isolate genomes ($n = 32$). Gene occurrence percentage >100% (that is, more than one copy per genome equivalent) in the GOS samples are represented as 100%. Gene descriptions are listed in Supplementary Table S2. GOS samples included in the comparison are listed in Supplementary Table S3.

As the GOS samples were passed through a 0.8-μM filter before sequencing (Rusch *et al.*, 2007), there is also poor representation of sequences from particle-associated cells.

Conclusions

Comparative genomic analysis of a bacterial lineage is a powerful approach for revealing ecological and evolutionary forces that influence genome content, and might form the basis for delineating ecologically differentiated clusters in nature. The substantial 16S rRNA sequence divergence within the roseobacters (11%; Buchan *et al.*, 2005), currently spanning a minimum of 45 described genera, makes this the broadest marine bacterial lineage for which a comparative genomic analysis has yet been undertaken. This taxonomic level is consistent, however, with current methodological resolution in microbial ecology, including target groups for 16S rRNA probes and primers (Alonso-Sáez *et al.*, 2007; Lami *et al.*, 2009), and efforts to assign taxon-specific biogeochemical roles (Alonso and Pernthaler, 2006; Mou *et al.*, 2008; Poretsky *et al.*, 2010).

The three predictive frameworks examined here for the Roseobacter genomes have previously been shown to correlate with the genome content in bacterial taxa, including phylogenetic relatedness in *Prochlorococcus* (Garczarek *et al.*, 2000; Bibby *et al.*, 2003; Rocap *et al.*, 2003), environmental resource partitioning in Vibrionaceae (a lineage with similar 16S rRNA divergence as the roseobacters; Hunt *et al.*, 2008), and trophic strategies in bacterial endosymbionts and aquatic bacterioplankton (Moran and Baumann, 2000; Lauro *et al.*, 2009). For the Roseobacter lineage, whole-genome content analysis of the 32 genomes produced 23 genome clusters (Figure 3) representing 20 unique combinations of clade, trophic strategy and environmental source. New sequences of Roseobacter strains may well increase the number of known genome clusters, particularly because two environmentally abundant 16S rRNA clades do not yet have reference genome sequences. While all three frameworks had statistically significant predictive power, none emerged as the potential overriding force imprinting Roseobacter genome content. Although other possible explanatory frameworks might have been considered here, all but two of the 23 genome clusters have unique clade–trophy–environment assignments (Figure 3), suggesting that these three frameworks together acceptably classify most of the variability in genome content.

The finding that trophic strategy correlates better than phylogeny or environment with Roseobacter gene inventories (ANOSIM, $R = 0.545$ vs 0.410 vs 0.296) was not anticipated at the outset of our analysis, at least in part because it is not a correlate that has been widely examined for marine

bacterial genomes. Nevertheless, the past decade has uncovered remarkable flexibility in the trophic strategies of marine bacterioplankton, suggesting that acquisition of alternate mechanisms for obtaining carbon and energy may be a strong evolutionary force in the ocean. The occurrence of several distinct trophic schemes within the taxonomically broad Roseobacter lineage provided an ideal opportunity to explore whether a bacterium's strategy for obtaining carbon and energy predicts other aspects of genome content. Differences in gene content among trophic groups were unfortunately dominated by hypothetical proteins, which provide little biological insight, although C1 and aromatic carbon oxidation genes and amino-acid transport and metabolism genes contributed to the signal. This concept of lifestyle imprinting of genome content, which has been explored in great detail for bacterial endosymbionts (for example, Moran and Baumann, 2000), may therefore also be important for understanding gene inventories of ocean microbes.

Roseobacter-like genes in the GOS data set showed significant variation in frequency across ocean basins, although only a fraction of all possible genes and gene systems appear to be shaped at this grand scale (Figure 6a). The GOS data set was also valuable for determining how well the genomes from the cultured roseobacters represent the repertoire and stoichiometry of genes in ocean-dwelling 'wild' roseobacters, an important perspective for assessing the relevance of this isolate-based genome analysis. Although the mismatch in frequency of some examined genes between isolates and the GOS data set suggests that the currently cultured strains may not yet provide a faithful representation of the prevalent natural Roseobacter populations, many genes and gene systems were indeed present at comparable frequencies (Figure 6b).

Overall, our analysis has firmly established roseobacters as ecological generalists, harboring large gene inventories and a remarkable suite of mechanisms by which to obtain carbon and energy. Further, this comparative analysis has illustrated that members of the lineage cannot be easily condensed into a few ecologically differentiated clusters; rather, each genome is largely unique in its assortment of genes for acquisition and transformation of carbon and nutrients. The fact that the best framework for predicting genome content is lifestyle strategy, not phylogeny, indicates that horizontal gene transfer and homologous recombination may be particularly dominant evolutionary forces in this marine bacterial lineage (possibly facilitated by an unusual gene transfer agent system that is prevalent; Biers *et al.*, 2008; Zhao *et al.*, 2009). Further insights into correlates of genome content, coupled with continued efforts to identify Roseobacter genes that are common in the world oceans, will better elucidate the functional roles of roseobacters in marine ecosystems.

Acknowledgements

This project was supported by grants from the Gordon and Betty Moore Foundation and the National Science Foundation (OCE0724017 and MCB0702125). We thank Dr S Santos for providing a list of universal genes.

References

Alonso C, Pernthaler J. (2006). Concentration-dependent patterns of leucine incorporation by coastal picoplankton. *Appl Environ Microbiol* **72**: 2141–2147.

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

Bibby TS, Mary I, Nield J, Partensky F, Barber J. (2003). Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. *Nature* **424**: 1051–1054.

Biers EJ, Wang K, Pennington C, Belas R, Chen F, Moran MA. (2008). Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl Environ Microbiol* **74**: 2933–2939.

Blankenship RE, Madigan MT, Bauer CE. (1995). *Anoxygenic Photosynthetic Bacteria*. Kluwer Academic Publishers: Dordrecht, Boston.

Brinkhoff T, Giebel HA, Simon M. (2008). Diversity, ecology, and genomics of the Roseobacter clade: a short overview. *Arch Microbiol* **189**: 531–539.

Buchan A, González JM, Moran MA. (2005). Overview of the marine Roseobacter lineage. *Appl Environ Microbiol* **71**: 5665–5677.

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

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. (2005). Algae acquire vitamin B_{12} through a symbiotic relationship with bacteria. *Nature* **438**: 90–93.

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP *et al.* (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90. .1-R90.16.

Eisen JA. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* **41**: 1105–1123.

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

Garczarek L, Hess WR, Holtzendorff J, van der Staay GWM, Partensky F. (2000). Multiplication of antenna genes as a major adaptation to low light in a marine prokaryote. *Proc Natl Acad Sci USA* **97**: 4098–4101.

González JM, Moran MA. (1997). Numerical dominance of a group of marine bacteria in the alpha-subclass of the class Proteobacteria in coastal seawater. *Appl Environ Microbiol* **63**: 4237–4242.

Goris J, Konstantinos KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**: 81–91.

Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, Lamerdin J *et al.* (2001). The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. *Photosynth Res* **70**: 53–71.

Howard EC, Sun SL, Biers EJ, Moran MA. (2008). Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environ Microbiol* **10**: 2397–2410.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.

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.

Karl DM. (2002). Nutrient dynamics in the deep blue sea. *Trends Microbiol* **10**: 410–418.

King GA. (2003). Molecular and culture-based analyses of aerobic carbon monoxide oxidizer diversity. *Appl Environ Microbiol* **69**: 7257–7265.

Lami R, Ghiglione J-F, Desdevises Y, West NJ, Lebaron P. (2009). Annual patterns of presence and activity of marine bacteria monitored by 16S rDNA–16S rRNA fingerprints in the coastal NW Mediterranean Sea. *Aquat Microb Ecol* **54**: 199–210.

Lauro FM, McDouglad D, Thomas T, Williams TJ, Egan S, Rice S *et al.* (2009). The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* **106**: 15527–15533.

Legendre P, Legendre L. (1998). *Numerical Ecology*. Second English edn. Elsevier Science, BV, Elsevier: Amsterdam, Netherlands.

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

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D *et al.* (2008). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**: D534–D538.

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

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

Moran MA, Belas R, Schell MA, González JM, Sun F, Sun S *et al.* (2007). Ecological genomics of marine roseobacters. *Appl Environ Microbiol* **73**: 4559–4569.

Moran MA, Buchan A, González JM, Heidelberg JF, Whitman WB, Kiene RP *et al.* (2004). Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432**: 910–913.

Moran NA, Baumann P. (2000). Bacterial endosymbionts in animals. *Curr Opin Microbiol* **3**: 270–275.

Mou X, Sun S, Edwards RA, Hodson RE, Moran MA. (2008). Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451**: 708–711.

Palenik B, Ren QH, Dupont CL, Myers GS, Heidelberg JF, Badger JH *et al.* (2006). Genome sequence of *Synechococcus* CC9311: insights into adaptation to a coastal environment. *Proc Natl Acad Sci USA* **103**: 13555–13559.

Polz MF, Hunt DE, Preheim SP, Weinreich DM. (2006). Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Phil Trans R Soc B* **361**: 2009–2021.

Poretsky RS, Sun S, Mou X, Moran MA. (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol*. Advance online publication doi:10.1111/j.1462–2920.2009.02102.x.

Rabouille S, Edwards CA, Zehr JP. (2007). Modeling the vertical distribution of *Prochlorococcus* and *Synechococcus* in the North Pacific Subtropical Ocean. *Environ Microbiol* **9**: 2588–2602.

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

Rusch DB, Halpern A, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: 398–431.

Santos SR, Ochman H. (2004). Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol* **6**: 754–759.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Bio Rev* **73**: 249–299.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: 394–397.

Shiba T, Simidu U, Taga N. (1979). Distribution of aerobic-bacteria which contain bacteriochlorophyll-*a*. *Appl Environ Microbiol* **38**: 43–45.

Shimada K. (1995). Aerobic anoxygenic phototrophs. In: Blankenship RE, Madigan MT, Bauer CE (eds). *Anoxygenic Photosynthetic Bacteria*. Vol. 2. Springer Netherlands, Dordrecht; Boston: Kluwer Academic Publishers, pp 105–122.

Snel B, Bork P, Huynen MA. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* **12**: 17–25.

Sorokin DY, Banciu H, van Loosdrecht M, Kuenen JG. (2003). Growth physiology and competitive interaction of obligately chemolithoautotrophic, haloalkaliphilic, sulfur-oxidizing bacteria from soda lakes. *Extremophiles* **7**: 195–203.

Stamatakis A, Hoover P, Rougemont J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* **57**: 758–771.

Swingley WD, Sadekar S, Mastrian SD, Matthies HJ, Hao J, Ramos H *et al.* (2007). The complete genome sequence of *Roseobacter denitrificans* reveals a mixotrophic rather than photosynthetic metabolism. *J Bacteriol* **189**: 683–690.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 1–14.

Wagner-Döbler I, Biebl H. (2006). Environmental biology of the marine Roseobacter lineage. *Annu Rev Microbiol* **60**: 255–280.

Wagner-Döbler I, Ballhausen B, Berger M, Brinkhoff T, Buchholz I, Bunk B *et al.* (2009). The complete genome sequence of the algal symbiont *Dinoroseobacter shibae*: a hitchiker's guide to life in the sea. *ISME J* **4**: 61–77.

West NJ, Obernosterer I, Zemb O, Lebaron P. (2008). Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean. *Environ Microbiol* **10**: 738–756.

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

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**: 432–466.

Zhao YL, Wang K, Budinoff C, Buchan A, Lang A, Jiao NZ *et al.* (2009). Gene transfer agent (GTA) genes reveal diverse and dynamic Roseobacter and Rhodobacter populations in the Chesapeake Bay. *ISME J* **3**: 364–373.

Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D *et al.* (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* **10**: 147–161.

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