The ISME Journal (2009) 3, 352–363
© 2009 International Society for Microbial Ecology All rights reserved 1751-7362/09 $32.00
www.nature.com/ismej

ORIGINAL ARTICLE

Distribution of microbial terpenoid lipid cyclases in the global ocean metagenome

Ann Pearson[1] and Douglas B Rusch[2]
[1]*Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA and* [2]*J Craig Venter Institute, Rockville, MD, USA*

The bacterial terpenoid lipids known as hopanoids are fundamental tools for interpreting ancient microbial communities. Their degradation products, the hopanes, are found in sedimentary rocks throughout the geologic record. These compounds are presumed to be analogous to the sterols of eukaryotes, yet although the eukaryotic requirement for sterols is universal, hopanoid biosynthetic capacity is not ubiquitous among marine bacteria. Among the 9.8 million shotgun reads from the Sorcerer II Global Ocean Sampling (GOS) expedition, 148 contain putative coding sequence for bacterial squalene-hopene cyclases (SHCs). SHCs encoded by α-Proteobacteria potentially related to Rhodospirillaceae dominate these hits, especially in the open ocean and in tropical regions. Planctomycetes and β-Proteobacteria contribute more SHC-encoding sequences, and therefore presumably more hopanoid production, to coastal and temperate environments. Although sequences nominally related to α- and β-Proteobacteria outnumber other taxa in marine and coastal environments, there is large phylogenetic distance between GOS sequences and known species. Assuming that the environments sampled here are broadly representative of a wide range of surface ocean climates, depositional settings and temporal periods, the data suggest a fundamental function for Proteobacteria in the development of the geologic record of hopanes.

The ISME Journal (2009) **3,** 352–363; doi:10.1038/ismej.2008.116; published online 27 November 2008
Subject Category: geomicrobiology and microbial contributions to geochemical cycles
Keywords: biomarkers; lipid biosynthesis; metagenomics

Introduction

Geologists interested in the biological and environmental history of Earth have a particular interest in the synthesis of polycyclic terpenoid lipids. These compounds include tetracyclic sterols (for example, cholesterol) and pentacyclic hopanoids (for example, bacteriohopanetetrol; Forster *et al.*, 1973; Rohmer and Ourisson, 1976), as well as less common biomolecules such as tetrahymanol (Mallory *et al.*, 1968) and malabaricatriene (Behrens *et al.*, 1999). All of these molecular classes are synthesized from the acyclic isoprenoid precursor, squalene (Woodward and Bloch, 1953), by a set of homologous enzymes known as polycyclic terpenoid cyclases. Such enzymes include squalene-hopene cyclase (SHC; Wendt *et al.*, 1997) and steroidal oxidosqualene cyclase (OSC; Thoma *et al.*, 2004). Although sterols are universal biosynthetic or dietary requirements among eukaryotes, the hopanoids of prokaryotes are irregularly distributed phylogenetically (Rohmer *et al.*, 1984).

Terpenoid lipid products, particularly hopanes and steranes, are detectable in the record of sedimentary rocks over timescales of billions of years (Ga). Because hopanoid lipids display regional and temporal patterns in their distributions, there must be taxonomic and/or physiological signals preserved in the records of these compounds. For example, 3-methylhopanes were reported at 1.64 Ga from the Barney Creek formation of Australia (Brocks *et al.*, 2005). These compounds were interpreted to record an abundance of type I aerobic, methanotrophic bacteria in the upper water column of the Mesoproterozoic ocean at this time. Reports of 2-methylhopanes from even more ancient strata (2.7 Ga Pilbara Craton, Australia; 2.45 Ga Matienda Formation, Canada) have been interpreted to indicate the presence of Cyanobacteria and therefore the very ancient evolution of oxygenic photosynthesis (Brocks *et al.*, 1999; Summons *et al.*, 1999; Dutkiewicz *et al.*, 2006) although the syngeneity of these molecules recently has been questioned (Rasmussen *et al.*, 2008).

The desire to map lipids to their sources is generating diverse approaches to the new field of microbial 'lipidomics' (for example, Hinrichs *et al.*, 1999; Pearson *et al.*, 2003, 2007; Coolen *et al.*, 2004,

Correspondence: A Pearson, Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, MA 2138, USA.
E-mail: pearson@eps.harvard.edu

Received 17 September 2008; accepted 27 October 2008; published online 27 November 2008

2007; Damste *et al.*, 2004a; Thiel *et al.*, 2007). In particular, the realization that >99.9% of known microbial diversity remains uncultured (Hugenholtz and Pace, 1996) prompts many questions. What lipids do these unstudied organisms make? How do we link lipids and unknown taxa with environmental conditions, given the vastness of microbial diversity? To understand the geobiological record, we must understand the distribution of hopanoid producers in the context of these questions.

Metagenomics offers one culture-independent option. Bioinformatics approaches have been used previously to identify terpenoid cyclase genes both in pure cultures (Perzl *et al.*, 1997; Tippelt *et al.*, 1998; Bode *et al.*, 2003; Pearson *et al.*, 2003) as well as in environmental samples (Pearson *et al.*, 2007, 2008). SHCs can be identified readily based on their numerous conserved amino acids (AAs; Hoshino and Sato, 2002; Summons *et al.*, 2006; Fischer and Pearson, 2007). Specific functional motifs (Feil *et al.*, 1996; Wendt *et al.*, 2000) and analysis of secondary structures (Wendt *et al.*, 1997; Schulz-Gasch and Stahl, 2003) show that the two classes can be distinguished by sequence differences alone. There are no abiotic processes or other enzymes known to perform the functions of SHCs (Ourisson *et al.*, 1987; Wendt *et al.*, 2000). Thus, identifying the putative SHC-encoding genes from a representative set of marine samples should provide a snapshot of the potential of those communities to produce hopanoids.

The data collected as part of the Sorcerer II Global Ocean Sampling (GOS) expedition provide an opportunity to examine environmental SHCs. The public GOS data (Rusch *et al.*, 2007) include and extend the previously published results from the Sargasso Sea (Venter *et al.*, 2004). Data have been released from 77 samples representing 69 locations. Most of these samples reflect the bacterial size class (0.1–0.8 μm), although some larger particulate classes (0.8–3.0 μm) occasionally were included. The vast majority of the data are from small prokaryotes and include 9.8 million sequence fragments, representing 8.3 billion base pairs of DNA and containing more than 6 million open reading frames (ORFs) (Yooseph *et al.*, 2007).

Preliminary work with the original Sargasso Sea metagenomic data (Venter *et al.*, 2004) identified 23 complete sequences or fragments of putative SHCs and 2 fragments of OSCs (Pearson *et al.*, 2007). Here we augment those results with a comprehensive analysis of GOS sites 00(b), Sargasso Sea through GS149, West Coast Zanzibar, Tanzania.

Materials and methods

Data

The GOS metagenome has been described elsewhere (Rusch *et al.*, 2007). Briefly, microbial samples were collected as part of the Sorcerer II expedition between 8 August 2003 and 12 September 2005 (Supplementary Figure S1 and Table S1). Specimens were collected from surface (generally 1–3 m, but occasionally as deep as 30 m), primarily marine, environments at approximately 320-km intervals. Total DNA was extracted from one or more particle size fraction for every site (Rusch *et al.*, 2007), and random-insert clone libraries were constructed for shotgun sequencing. The public data release is accessible through Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA; http://camera.calit2.net/index.php).

Recruitments

A customized version of the Advanced Reference Viewer tool (DB Rusch, unpublished) was used to compare translated GOS sequences to a reference sequence consisting of an artificial concatenation of SHCs obtained from microbial genomes (92 sequences, average length 630 AA), each separated by a spacer string of 500 X characters. It was current as of February 2008, and is believed to represent a comprehensive picture of bacterial SHCs known to date. It includes most species and all known genera that contain a gene encoding for SHC. Selected SHCs were excluded in cases of near redundancy, for example, *Burkholderia mallei* was retained whereas *Burkholderia pseudomallei* was not. Overall there were <10 such exclusions. Criteria for identifying SHC sequences within bacterial genomes have been described previously (Tippelt *et al.*, 1998; Hoshino and Sato, 2002; Fischer and Pearson, 2007). The recruitment was performed using NCBI tBLASTn (Altschul *et al.*, 1997) (nondefault parameters: -F 'm L' -U T -e 1e-10 -z 3000000000 -b 10000 -K 10000) and the alignments were visualized through the Advanced Reference Viewer. Aligned reads were filtered at ⩾40% AA local alignment identity and the minimum recruited fragment length was set to ⩾75 AA.

Reciprocal BLAST and phylogenetic binning

The AA translation for each putative SHC encoding ORF was examined by reciprocal BLAST (Altschul *et al.*, 1997). DNA sequences were translated into ORFs using the NCBI ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) and were checked for homologous sequence preceding the nominally identified start codon using the ExPASy Translate Tool (http://www.expasy.ch/tools/dna.html). Each edited amino-acid sequence was then matched with its closest relative among all genomic data in the public domain using the nr/nt nucleotide database and the NCBI (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) implementation of tBLASTn. Typical expectation values for 75-mer fragments compared to sequences of known SHCs are $<10^{-5}$; *E*-values are much smaller (for example, $E < 10^{-50}$) for 200-mer fragments. Sequences were separated into putative SHCs and putative OSCs (Supplementary Tables S2 and S3).

Phylogenetic analysis
Amino-acid alignments were performed using Clustal X (Thompson *et al.*, 1997). A guide alignment was prepared from full-length genomic reference sequences, and the GOS fragments subsequently were aligned to the guide. Multiple alignment parameters were gap penalty 12.0, gap extension 0.10, BLOSUM weight matrix. After alignment, incomplete gene segments were masked at their N-terminal and C-terminal ends to compensate for length heterogeneity before submitting to the PHYML Online Server (http://atgc.lirmm.fr/phyml/). Parameters for likelihood analysis were 100 bootstrap replicates, WAG substitution model, starting tree BIONJ, and optimized topology and branch lengths (Guindon and Gascuel, 2003). The SHC sequence of *Alicyclobacillus acidocaldarius* served as the outgroup.

Results and discussion

Recruitment
The GOS data currently in the public domain span stations from the Sargasso Sea, north to the Canadian Atlantic Margin, south through the Caribbean, across the Panama Canal into the Eastern Tropical North Pacific, south to the equatorial Galapagos Islands, across the open Pacific Ocean to Tahiti, and through the Indian Ocean and south to South Africa. The range of climates, geographic locations and local environments is broad. Yet within the data there is sufficient density of samples to allow some clustering into type categories such as ‘open ocean’, or ‘coastal’ (Supplementary Table S1). Most individual sample locations have >100 Mb of available DNA sequence, and some have considerably more (Rusch *et al.*, 2007). A total of 9.8 million shotgun reads, representing ~8.3 Gbp of DNA sequence (or 4150 whole-genome equivalents, if each microbe contains 2 Mb of DNA), were analyzed.

DNA fragments that would putatively encode for SHCs were identified from the total data, yielding 6490 sequence fragments, or an equivalent of 70 fragments recruited per genomic sequence on average (Figure 1). However, the density is not uniform and there are patterns of recruitment that show phylogenetic clustering (Figure 2). A large absolute number of fragments (~33% of the 6490) originate from GS33, the hypersaline lagoon of Floreana Island in the Galapagos (Figures 1 and 2).

Importantly, these recruited fragments do not represent 6490 different sequences. The same sequences are recruited by multiple species as a consequence of the sequence similarity between most SHCs. For example, the fragment showing the greatest identity to *Nitrosomonas europaea* (62%) is the same fragment (JCVI_READ_109114075901) that shows 63% identity to *Nitrosospira multiformis* and *Nitrococcus mobilis* but only 58% identity to *Methylococcus capsulatus*. This is counted as four recruitment events although it is only one sequence. There are many analogous cases, and among the 6490 fragments, ultimately there are just 148 and 22 different SHCs and OSCs, respectively (Supplementary Tables S2 and S3). All recruited sequences were subjected to verification by reciprocal tBLASTn. Although assignments of function based only on sequence must remain putative, the results suggest that if the total data represent 4150 microbial genome equivalents, 4% of all cells in the GOS samples contain an SHC or OSC and potentially could produce polycyclic terpenoids.

Each major phylogenetic group (Figures 2a–e) shows a different pattern of recruitment. SHCs of Proteobacteria (a–c) recruit a high density of fragments having moderate (50–60%) identity, although there are somewhat fewer for δ-Proteobacteria (c). In contrast, hits to Cyanobacteria, Actinomycetes and Acidobacteria (d, e) are much less abundant in this identity range. Below 50% identity, nearly all SHCs recruit a high density of fragments. This reflects the overall similarity within this protein family among all taxa except for the phylogenetically distant Planctomycetes. Planctomycetes recruit a low total density of fragments, although a few fragments show high individual identity.

Among all species, the α-Proteobacterium *Rhodospirillum rubrum* has both the greatest number and greatest identity of recruited sequences (Figure 2a). SHCs from locations in all ocean basins map to *R. rubrum* at high (60–70%) identity. A moderate density of fragments also recruits with high identity to *Acidiphilium cryptum* and to *Granulobacter bethesdensis*. These tend to be dominantly from the Galapagos hypersaline lagoon (GS33). Only one sequence in any GOS sample was identified as having as high as ~70% identity to a known species. Again it is an apparent relative of *R. rubrum*.

Among β-Proteobacteria, a number of sequences from GS33 and from the Sargasso Sea show ~60% to 65% identity to known organisms, and it appears that *Ralstonia* spp. are the most similar. There is also a sequence from Newport Harbor (GS08) that appears more closely related to ammonium oxidizers (nitrosifyers). This is the fragment identified above as a relative of *N. mobilis* or *N. multiformis*. It would be tenuous to classify it specifically as a nitrosifyer, however, as although a 60–65% sequence identity is shared among all known SHCs of β-Proteobacteria, the limited data for ammonium oxidizers suggests their SHCs are 70–75% identical. It is important to note again that all of the data from GS00(a) have been excluded. GS00(a) contains a

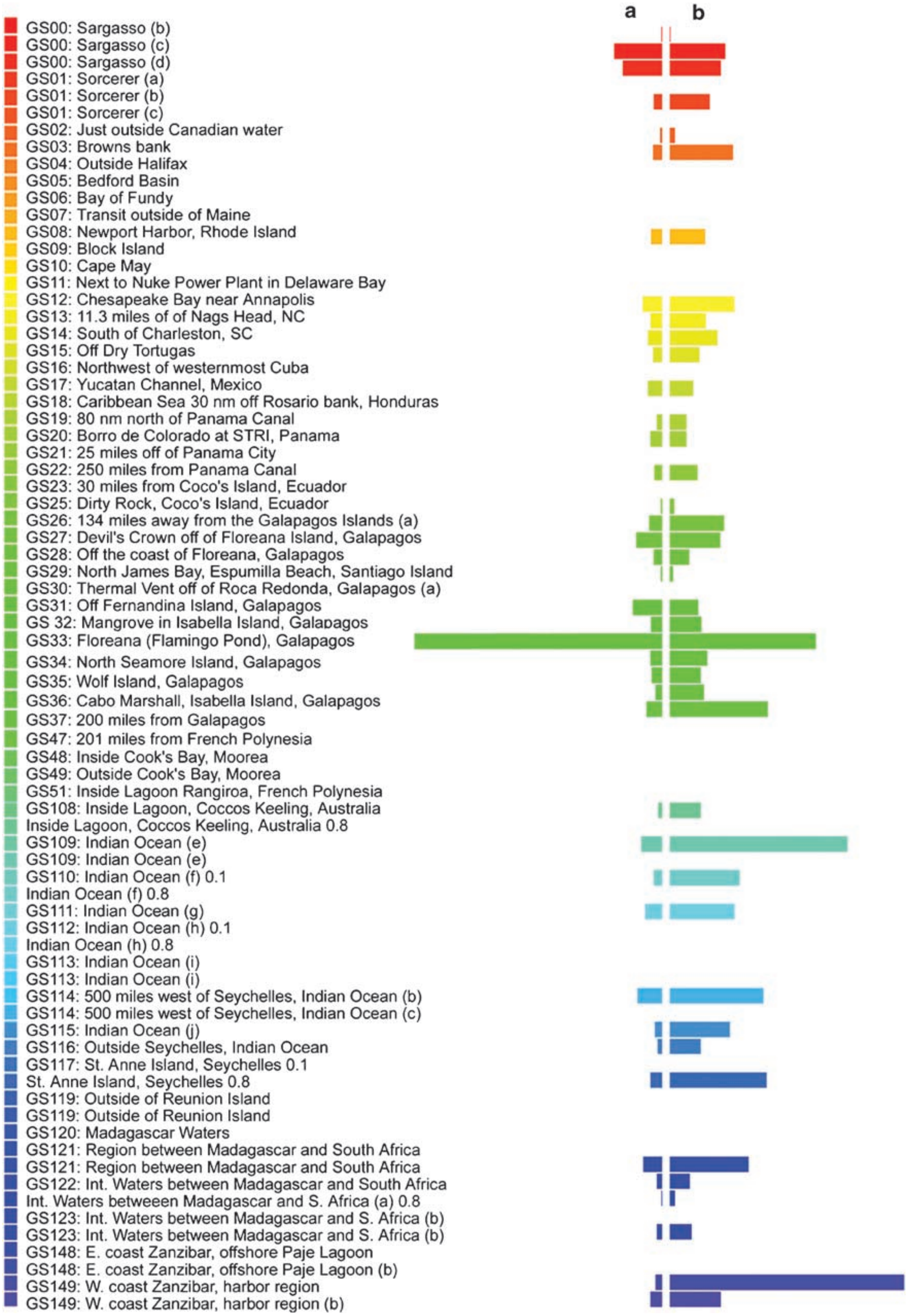

Figure 1 Abundance of terpenoid cyclase fragments identified using the CAMERA Reference Viewer. Scales of histograms reflect 6490 total recruited fragments (**a**) and the normalized distribution of these fragments (**b**). Colors identify Global Ocean Sampling (GOS) stations and match the colors in Figure 2.

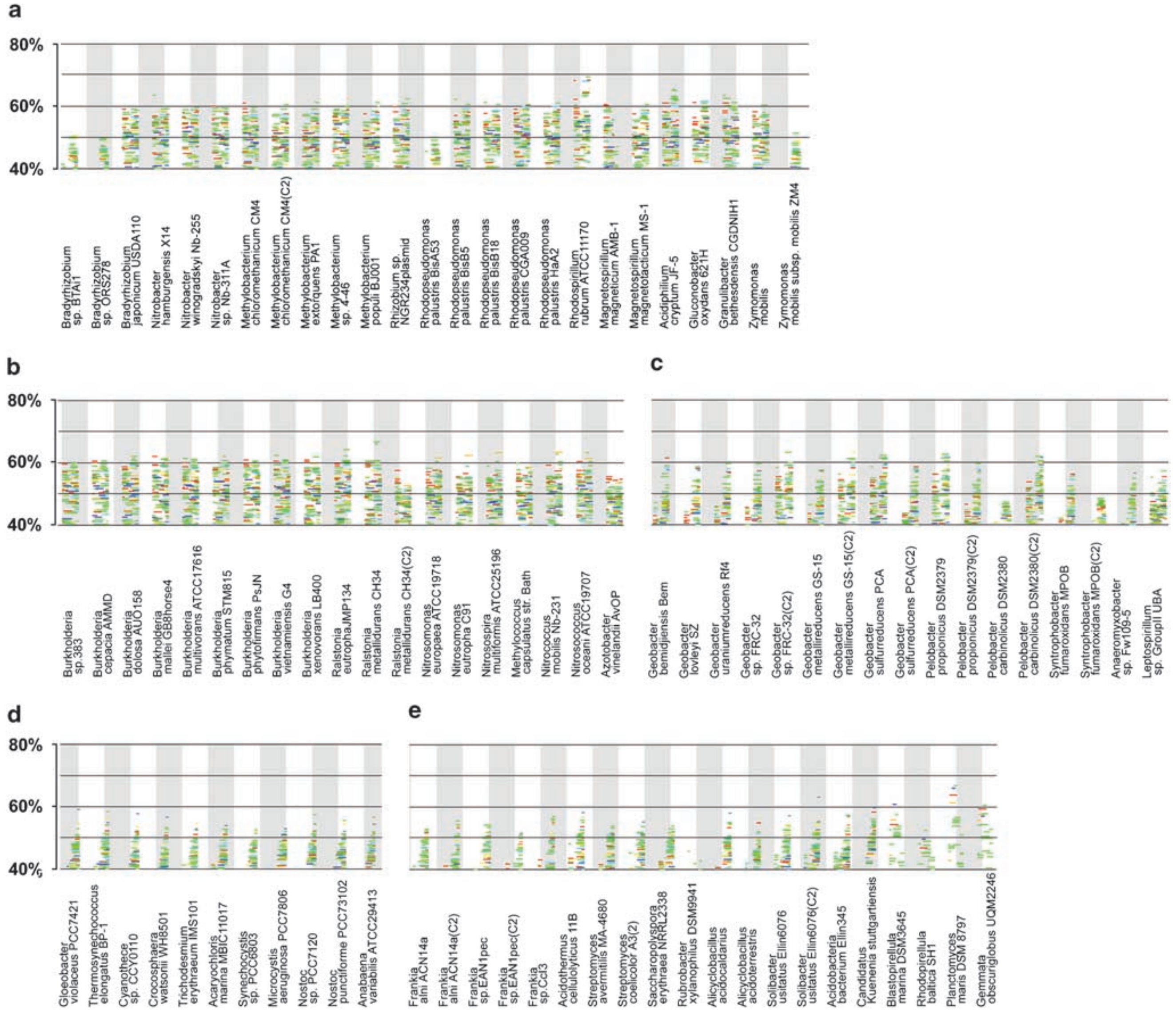

Figure 2 Recruitment maps. The *x* axis of each panel corresponds to concatenated AA sequences for squalene-hopene cyclases (SHCs), grouped by phylogeny. The *y* axis indicates the sequence identity of an alignment between a Global Ocean Sampling (GOS) sequence and the reference SHC. The identity ranges from 80% (top) to 40% (bottom). Individual GOS sequencing reads are colored to reflect the sample from which they derive, as in Figure 1. Geographically proximal samples have similar colors. (**a**) α-Proteobacteria; (**b**) β- and γ-Proteobacteria; (**c**) δ-Proteobacteria and *Leptospirillum* sp.; (**d**) Cyanobacteria; (**e**) others.

nearly closed genome (Venter *et al.*, 2004) of a strain very similar to *Burkholderia* sp. 383 (Stanier *et al.*, 1966), and it is a suspected contaminant in samples from that site (DeLong and Karl, 2005; Mahenthiralingam *et al.*, 2006). This analysis is concerned only with GS00(b)-GS149.

Among δ-Proteobacteria, sequences with 60–70% identity to Geobacteraceae can be detected in the Galapagos lagoon and in sites GS11-14, the Sargasso Sea, and the Indian Ocean. Known Geobacteraceae share SHC sequence identity of only 58%, but when split into 'copy 1' and 'copy 2' designations, they have 70% identity within each of the subsets. Although it is unlikely that members of the anaerobic genus *Geobacter* are important contributors to the GOS data, these hits are consistent with overall placement within the Geobacteraceae. A lower fragment density is recruited to *Syntrophobacter*, *Anaeromyxobacter*, and to the Nitrospirales genus *Leptospirillum* (whose SHC is similar to SHCs of δ-Proteobacteria and is thus included in Figure 2c).

The recruitment to known SHCs of Cyanobacteria (Figure 2d) yields no sequences that exceed 60% identity among any of the locations. For reference, 55–60% identity is the approximate threshold range for distinguishing between SHCs from different major bacterial groups, for example, Cyanobacteria from α-Proteobacteria (Pearson *et al.*, 2007). There are two candidate cyanobacterial fragments between 55% and 60%: one is a fragment from GS33 that is ~100 AA long, has 58% identity to *Thermosynechococcus*, and is classified as such by reciprocal tBLASTn (Table 1; Supplementary Table S2). The other is a fragment from St. Anne Island, Seychelles, showing high local identity to the C-terminal end of *Gloeobacter violaceus*. However, the best overall match for this sequence (JCVI_READ_1105333626861) is to the Acidobacterium *Solibacter usitatus* (Figure 2e). No other fragments were classified by tBLASTn as belonging to Cyanobacteria, and in general there is a low fragment density throughout Figure 2d, particularly when compared to the fragment density observed for α-, β- and γ-Proteobacteria (Figures 2a and b).

There are a number of sequences that are similar to *Planctomyces maris*. These, like α-Proteobacteria, are detected across all ocean basins. They are also found in GS33. This suggests a widespread distribution of SHC homologs from Planctomycetes. Finally, there are very low identities among sequences recruited to the Actinomycetes and Acidobacteria, consistent with the fact that these soil bacteria are not expected to be found in marine samples.

Individual classification of recruited fragments

The 148 individual SHC fragments were classified according to potential phylotype by tBLASTn (Table 1; Figure 3). The proportions of phylotypes vary between open-ocean (Figure 3a) and coastal (Figure 3b) sequences, as well as GS33, the hypersaline lagoon at Floreana Island, Galapagos (Figure 3c). Other sample types such as fresh water, estuarine, mangrove or reef sites do not have enough sequence data to estimate proportional distribution. Importantly, the hits were not filtered for uniqueness and no attempt was made to condense them by operational taxonomic units (OTUs). They should represent the proportional abundances of SHC-containing species in each sample, assuming the sample collection methods and DNA extraction efficiencies accurately reflect the microbial community composition.

The results from tBLASTn are consistent with a dominance of Proteobacteria. Because SHCs of α-, β- and γ-Proteobacteria are so similar in sequence to each other, they all appear to have high proportional representation when viewed with the Advanced Reference Viewer (Figures 2a and b). However, when only the best homolog for each sequence is considered, it is apparent that most of the hits have greater identity to α-Proteobacteria and lesser to β-Proteobacteria. SHCs with best matches to γ-Proteobacteria are much less common (Figure 3).

The AA identities for SHC fragments assigned putatively to α-, β- and γ-Proteobacteria generally are 50–60% (Supplementary Table S2). Although these values represent considerable phylogenetic distances from known species, they are within the (low) range of assignable phylogeny at the major group level. This indicates the environmental sequences probably belong to novel families within these groups. In contrast, many of the putative SHCs that associate with Acidobacteria (Figure 3) have

Table 1 Phylogeny of SHC fragments across GOS sample types

	α	*β*	*γ*	*δ*	*Acidobacteria*	*Planctomycetes*	*Cyanobacteria*	*Total*
Open ocean (surface)	24	7	2	4	5	6	0	48
Coastal, upwelling and harbor	10	10	2	2	8	9	0	41
Estuary	1	0	0	2	0	0	0	3
Fresh water	0	2	1	0	0	0	0	3
Reefs	0	1	0	0	0	0	0	1
Hypersaline lagoon	16	4	1	7	10	8	1	47
Mangrove	0	0	0	3	0	2	0	5
								0
Total	51	24	6	18	23	25	1	148

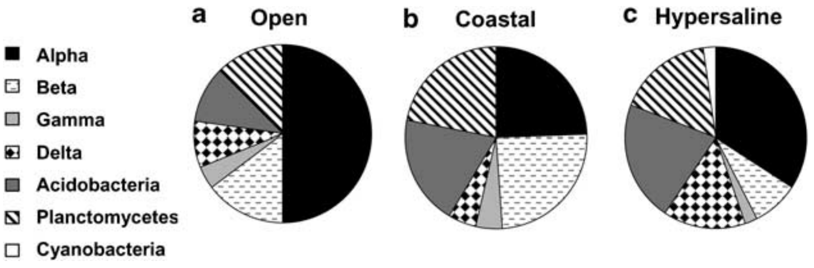

Figure 3 Phylogenetic affiliation of recruited squalene-hopene cyclase (SHC) fragments, as a function of sample type as designated in Supplementary Table S1: (**a**) open ocean, (**b**) coastal, harbor or upwelling, (**c**) Galapagos hypersaline lagoon, Floreana Island, GS33.

very poor identity to characterized Acidobacteria. The SHCs assigned nominally to Acidobacteria are only ∼40% identical to known Acidobacteria (Supplementary Table S2). This suggests that many of these hits could derive from uncultured major bacterial groups and not from true Acidobacteria. These results highlight the fact that tBLASTn cannot account for SHCs deriving from bacteria that have not yet been grown in culture (for example, SAR-86 or OP-11). All of the data have been forced to bin with their 'nearest' characterized relatives, regardless of how large that phylogenetic distance might be. Overall, however, the data are consistent with classification of these sequences as putative SHCs. The nearest sequence and structural homologs to SHCs are the other terpenoid cyclase enzymes, sporulene cyclase (Bosak *et al.*, 2008) and OSC (Thoma *et al.*, 2004). Assuming that the lower bound for identifying an SHC is the distance between these classes, SHCs and sporulene cyclases share an average 27% identity, whereas SHCs and OSCs share 19% identity.

There are clear differences between GOS samples from open-ocean sites and from coastal regions. In the open ocean, α-Proteobacteria account for 50% of all SHCs, whereas they account for ∼25% in coastal waters (Figure 3). Sequences from β-, γ- and δ-Proteobacteria contribute an additional ∼25% of all sequences to open-ocean sites, with the remaining ∼25% partitioned between Acidobacteria and Planctomycetes. The coastal sites show nearly equal abundances of α- and β-Proteobacteria, with again minor contributions from γ- and δ-Proteobacteria. Proportionally more coastal sequences are distributed among putative Acidobacteria and Planctomycetes.

The hypersaline lagoon from the Galapagos is a special case, and its SHC profile appears as a hybrid between the open-ocean distribution and the coastal type; although with a high ratio of α- to β-Proteobacteria, it is somewhat more similar to the open-ocean profile (Figure 3). It also contains the sole GOS fragment with optimal identity to Cyanobacteria.

Comparison to 16S rRNA distributions

The presence of a significant proportion of novel SHCs is consistent with the observed phylogenetic distance of 16S rRNA genes within the GOS samples. The GOS data contain 811 distinct ribotypes when classified at a 97% identity threshold, out of 9905 16S rRNA fragments sequenced from sites GS00-GS149 (Rusch *et al.*, 2007). Among those that could be classified, α-Proteobacteria, including SAR-11 (*Pelagibacter* spp.), were the most abundant (32%). A total of 60 dominant ribotypes accounted for 73% of all 16S sequences (Rusch *et al.*, 2007), but very few of these are represented by close relatives having completed genome sequencing projects, and 3.4% diverged by more than 10% from any publicly available rRNA gene (Rusch *et al.*, 2007). This suggests that at least 3.4% of the total DNA complement of GOS samples represents unknown bacterial families.

It is unknown how many of the putative SHC-encoding genes within the data could derive from these divergent lineages, but many (if not most) of them may, considering that none of the major oceanic phylotypes contains genes for SHC. The greatest density of total DNA sequence in the GOS data aligns with the genome of *Pelagibacter ubique* (SAR-11). This species does not contain SHC-encoding genes (Rappe *et al.*, 2002), nor do any known strains of the other abundant marine microbes, for example, *Prochlorococcus* spp. (Rappe *et al.*, 2002; Rocap *et al.*, 2003). Other dominant taxa among the 16S rRNA data whose relatives contain no known terpenoid cyclases are Bacteroidetes, ε-Proteobacteria and *Chloroflexi*. And although marine *Roseobacter* spp. remain understudied, none of the *Roseobacter* genomes recently sequenced as part of the Gordon & Betty Moore Foundation marine microbial genome sequencing project (http://www.jcvi.org/cms/research/projects/microgenome/overview/) contain sequences for SHCs. Other important ribotypes, including SAR86, have no cultivated close relatives at all. However, the numerous SAR86-affiliated scaffolds reveal no apparent instances of terpenoid cyclases (DB Rusch, unpublished). Further examination of scaffolded sequences may aid the challenge of assigning SHC-encoding genes to their source taxa. The greatest fractional abundance of 16S ribotypes in the GOS data are α-Proteobacteria, as are apparently the greatest proportion of SHC-encoding genes. Assigning the above to families and genera remains a significant challenge.

Given the sparse data, it also is difficult to determine if SHC-containing taxa have regional patterns of distribution. Previous analysis of the GOS 16S rRNA data showed there is regionality beyond simple sample classifications such as 'open ocean' or 'coastal' (Rusch *et al.*, 2007). Although SAR11 and SAR86 are ubiquitous, other phylotypes are distributed by geography and environment. The 16S rRNA data revealed six major habitats, as well as several other specific environments represented by a single location each (Rusch *et al.*, 2007). The biggest differences were observed between samples from tropical locations and samples from temperate locations. To see if SHCs showed similar geographic patterns, we prepared a spot-density plot (Figure 4) using the same categories (Rusch *et al.*, 2007), plus a new division (blue) to encompass all of the recently released data from the Indian Ocean. Unlike in the analogous 16S map, here there appear to be few (if any) patterns. This likely is a function of the much smaller numbers of SHC-encoding sequences relative to 16S genes per site, that is, there is inadequate sampling density.

This challenge is highlighted in the bottom row of Figure 4, which shows the number of hits in each

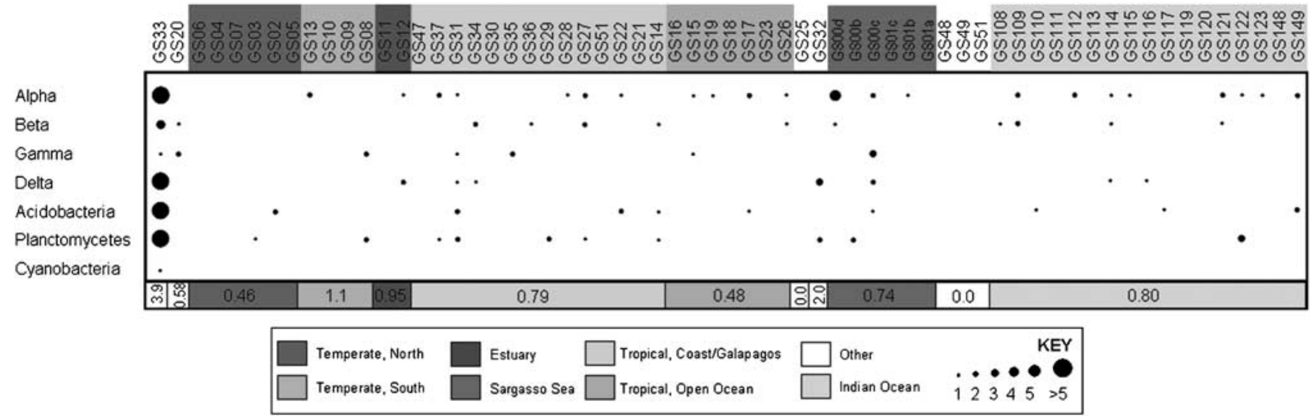

Figure 4 Squalene-hopene cyclase (SHC) fragments grouped geographically in the same order as the major 16S rRNA sample similarity clusters identified by Rusch *et al.* (2007). Numbers in the bottom row indicate the normalized abundance of SHCs in these geographic categories, defined as hits per 100 Mb of total sequence in that similarity cluster. Indian Ocean samples are shown separately, as they were not included in Rusch *et al.* (2007).

category as a function of the total amount of DNA sequence in that category. Site GS33 has the greatest density (3.9 hits per 100 Mb). Other regions with >1 hit per 100 Mb are the 'southern' temperate cluster including GS13 (Nag's Head) and GS08 (Newport Harbor, classified as southern by proximity to the Gulf Stream); and GS32, which is a mangrove estuary in the Galapagos. All other categories have <1.0 SHC hit per 100 Mb of DNA. Most individual GOS samples contain only 100–300 Mb of total shotgun sequence, and thus most locations have 0–5 instances of detectable SHC fragments. To view distributional patterns on a sample-by-sample basis, it would likely require 10 × the sequence data per location, that is, all samples would need to be sequenced at the depth of GS33. It is interesting, however, that the highest density of fragments appears to occur closest to land.

Phylogenetic analysis of α-Proteobacterial SHCs

Among the hits affiliated with α-Proteobacteria (Table 1; Supplementary Table S2), test AA alignments showed there was a region containing enough overlapping fragments for maximum likelihood analysis. Of the 51 total hits from α-Proteobacteria, 22 spanned the majority of an alignable 240-mer segment at the C-terminal end of SHC. The remaining sequences either were too short or overlapped the front or middle sections of the SHC protein and were thus excluded.

These α-Proteobacterial sequences fall into three groups (Figure 5). Group 1 consists of sequences from a wide geographic distribution, including both Atlantic and Pacific Oceans and tropical and temperate sites. These sequences form a cluster with *R. rubrum* as their closest relative, in agreement with the results from tBLASTn. A second group of sequences clusters between *R. rubrum* and the branch containing *Gluconobacter oxydans* and *A. cryptum*, although all of these GOS hits also classify with *R. rubrum* by tBLASTn assignment (Supplementary Table S2). The sites from which they derive are exclusively tropical and mostly open ocean. A third group, although nominally affiliated with α-Proteobacteria, forms a cluster that is basal to all known genera of α-Proteobacteria. All of these sequences derive from the Galapagos hypersaline lagoon (GS33). Three of them are also identical over a 100-AA span, and the other is 98% identical. They represent the only detected instance of replicate sequencing of what appears to be a single OTU.

None of the GOS data align with other known genera of α-Proteobacteria, including *Rhodopseudomonas*, *Bradyrhizobium*, *Nitrobacter*, *Methylobacterium*, *Magnetospirillum* or *Zymononas* spp. This suggests that most GOS hopanoid producers are relatives of Rhodospirillaceae; and, if not, are related to unidentified families of α-Proteobacteria or are members of other unclassified groups of Proteobacteria.

Implications for the sedimentary record of marine hopanoids

To date, the sources of geohopanes have been identified based on correlation with the distribution of hopanoids in cultured bacteria (for example, Rohmer *et al.*, 1984; Zundel and Rohmer, 1985a; Summons *et al.*, 1999; Talbot *et al.*, 2001), or based on the presence of unique isotopic signatures harbored by selected compounds (Elvert *et al.*, 2000; Pancost *et al.*, 2000; Hinrichs, 2001; Thiel *et al.*, 2001). Examples of the latter include numerous instances of ^{13}C-depleted 3-methylhopanes, which are attributed to type I methanotrophs of the γ-Proteobacteria (Zundel and Rohmer, 1985a, b). Oxidation of methane provides a ^{13}C-depleted carbon source for these species, explaining the appearance of this signature in their hopanoids.

The correlation between hopanoids and their source organisms is more difficult when distinctive isotopic signatures are not present. In recent years much effort has been devoted to understand the

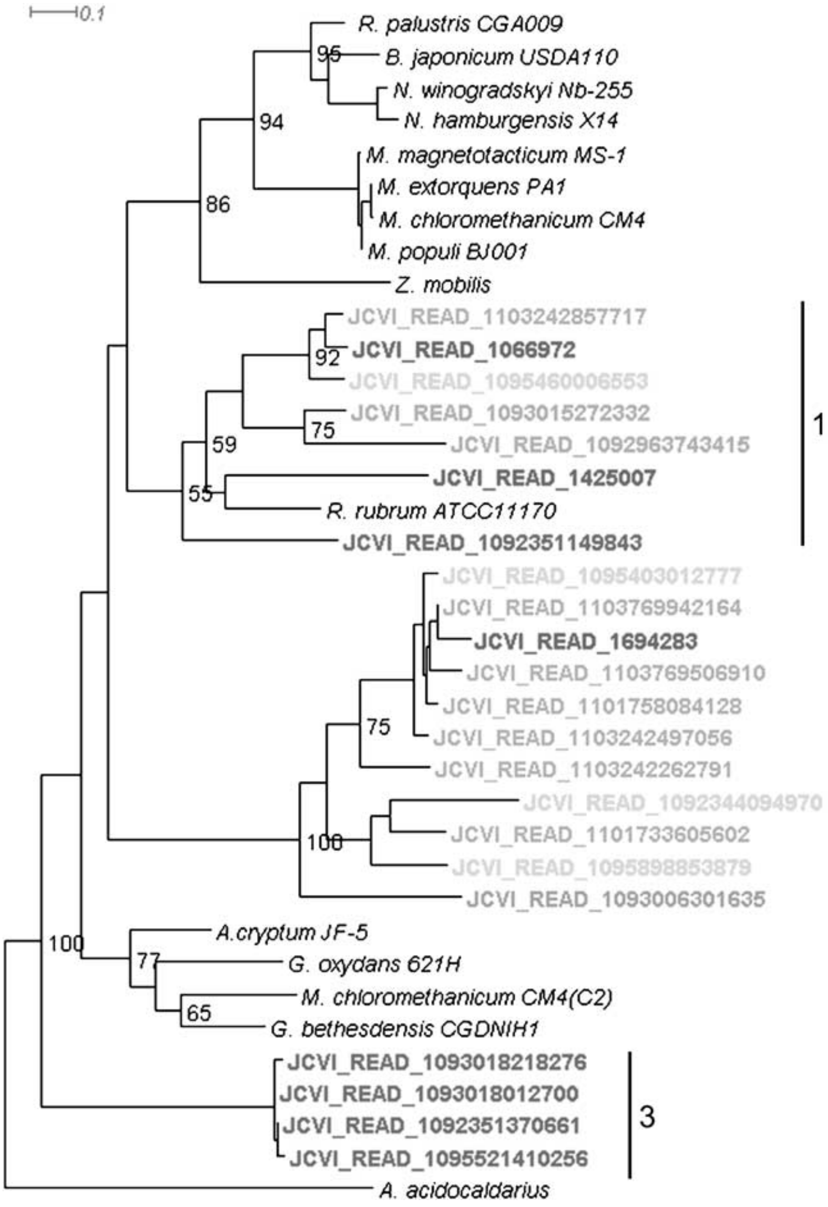

Figure 5 Maximum likelihood phylogenetic tree of partial squalene-hopene cyclase (SHC) sequences of α-Proteobacteria, as described in the text. Tree is based on alignment at the C-terminal end of SHC. Bootstrap values are out of 100 (only those with values ⩾50 are shown); branches are color coded to match geographic categories as in Figure 4.

distribution of unique polar head groups of bacteriohopanepolyols (Talbot *et al.*, 2003a, b, 2007a, b; Bednarczyk *et al.*, 2005; Sugden *et al.*, 2005; Talbot *et al.*, 2008). This method is quite promising for interpretations of modern water columns and sediments (Talbot *et al.*, 2003c; Wakeham *et al.*, 2007), but it has limited translation to the geologic record, as much of the information associated with these polar groups is lost through diagenesis.

The GOS data suggest that α-Proteobacteria could be important sources of hopanes in the sediment record, and that Proteobacteria in general produce the majority of hopanoids. In contrast, although the absence of cyanobacterial SHC sequences in the GOS samples does not mean that cyanobacterial producers of hopanoids are absent from the present-day ocean (they surely are present, for example, *Crocosphaera watsonii* and *Trichodesmium erythraeum*; Capone *et al.*, 1997; Zehr *et al.*, 2001, 2007), their proportional density among total hopanoid producers appears to be low.

Determining hopanoid profiles thus remains a significant challenge for species for which we have no cultured representatives. The low sequence identity of the SHCs in the GOS data indicates they represent species that have not been characterized in terms of their specific hopanoid compositions: most appear to represent novel genera or families. Early studies of cultured bacteria suggested that ~50% produced hopanoids (Rohmer *et al.*, 1984).

The GOS data show that the ability to make hopanoids is more restricted (⩽4% of total cells) and primarily resides with uncharacterized taxa. Establishing the phylogenetic identity of these putative sources of hopanoids is complicated by this rarity, which makes it difficult to obtain long, contiguous DNA sequence (contigs) from assembly of shotgun sequence data.

Any random sequencing approach thus will yield limited assembly data for SHC-encoding genes. This argues for co-reliance on other proxies such as stable isotope compositions, as well as for new metagenomic sequencing projects that have the ability to explore samples with greater depth and throughput (Sogin *et al.*, 2006). Other approaches also show some promise. Despite the limitations of applying PCR-based methods to functional genes that span wide phylogenetic distances, direct amplification of environmental DNA has yielded greater numbers and broader diversity of SHC-encoding genes than detected in the GOS data (Pearson *et al.*, 2007, 2008). These degenerate PCR methods, however, are accompanied by an unknown degree of bias. Thus, directed sequencing and random sequencing are two complementary approaches to expanding the known distribution of SHCs. Future work targeting genomic sequencing at the single-cell level also promises to yield more information (Zhang *et al.*, 2006; Stepanauskas and Sieracki, 2007). Together, these approaches will provide a more thorough description of microbial terpenoid producers and may help to interpret the geologic record. Finally, the environments analyzed here are conspicuously devoid of anaerobic habitats, and although the GOS study spans a wide diversity of latitude and longitude, the data are exclusive to the ocean's surface, ignoring its vast depth and sediments. The latter is of special interest, given the known production of terpenoid lipids by bacterial cultures and sedimentary consortia of anaerobes (Damste *et al.*, 2004b; Fischer *et al.*, 2005; Hartner *et al.*, 2005; Blumenberg *et al.*, 2006).

Acknowledgements

We thank Jan DeLeeuw for coining the term 'lipidomics' and members of the Pearson lab for helpful discussions. We acknowledge the Department of Energy (DOE), Office of Science, and Office of Biological and Environmental Research (DE-FG02-02ER63453), the Gordon and Betty Moore Foundation, and the J Craig Venter Science Foundation for funding to undertake this study. We gratefully acknowledge the governments of the countries of origin for the GOS samples: Canada, USA, United Kingdom, Mexico, Honduras, Panama, Costa Rica, Ecuador, French Polynesia, Australia, Seychelles, Madagascar and Tanzania. Genomic sequences from NCBI are available in the public domain courtesy of JGI (http://www.jgi.doe.gov/) and the J Craig Venter Institute. Special thanks are due to Marv Frasier, Saul Kravitz and the CAMERA team for their suggestions in support of the Advanced Reference Viewer. This work was supported by grants from the NSF and David and Lucille Packard Foundation (to AP).

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Bednarczyk A, Hernandez TC, Schaeffer P, Adam P, Talbot HM, Farrimond P *et al.* (2005). 32,35-Anhydrobacteriohopanetetrol: an unusual bacteriohopanepolyol widespread in recent and past environments. *Org Geochem* **36**: 673–677.

Behrens A, Schaeffer P, Bernasconi S, Albrecht P. (1999). 17(E)-13 alpha(H)-Malabarica-14(27), 17,21-triene, an unexpected tricyclic hydrocarbon in sediments. **30**: 379–383.

Blumenberg M, Kruger M, Nauhaus K, Talbot HM, Oppermann BI, Seifert R *et al.* (2006). Biosynthesis of hopanoids by sulfate-reducing bacteria (genus *Desulfovibrio*). *Environ Microbiol* **8**: 1220–1227.

Bode HB, Zeggel B, Silakowski B, Wenzel SC, Reichenbach H, Muller R. (2003). Steroid biosynthesis in prokaryotes: identification of myxobacterial steroids and cloning of the first bacterial 2,3(*S*)-oxidosqualene cyclase from the myxobacterium *Stigmatella aurantiaca*. *Mol Microbiol* **47**: 471–481.

Bosak T, Losick RM, Pearson A (2008). A polycyclic terpenoid that alleviates oxidative stress. *Proc Natl Acad Sci USA* **105**: 6725–6729.

Brocks JJ, Logan GA, Buick R, Summons RE. (1999). Archean molecular fossils and the early rise of eukaryotes. *Science* **285**: 1033–1036.

Brocks JJ, Love GD, Summons RE, Knoll AH, Logan GA, Bowden SA. (2005). Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437**: 866–870.

Capone DG, Zehr JP, Paerl HW, Bergman B, Carpenter EJ. (1997). Trichodesmium, a globally significant marine cyanobacterium. *Science* **276**: 1221–1229.

Coolen MJL, Muyzer G, Rijpstra WIC, Schouten S, Volkman JK, Damste JSS. (2004). Combined DNA and lipid analyses of sediments reveal changes in Holocene haptophyte and diatom populations in an Antarctic lake. *Earth Planetary Sci Lett* **223**: 225–239.

Coolen MJL, Volkman JK, Abbas B, Muyzer G, Schouten S, Damste JSS. (2007). Identification of organic matter sources in sulfidic late Holocene Antarctic fjord sediments from fossil rDNA sequence analysis. *Paleoceanography* **22**: PA2211, 10.1029/2006PA001309.

Damste JSS, Muyzer G, Abbas B, Rampen SW, Masse G, Allard WG *et al.* (2004a). The rise of the rhizosolenid diatoms. *Science* **304**: 584–587.

Damste JSS, Rijpstra WIC, Schouten S, Fuerst JA, Jetten MSM, Strous M. (2004b). The occurrence of hopanoids in planctomycetes: implications for the sedimentary biomarker record. *Org Geochem* **35**: 561–566.

DeLong EF, Karl DM. (2005). Genomic perspectives in microbial oceanography. *Nature* **437**: 336–342.

Dutkiewicz A, Volk H, George SC, Ridley J, Buick R. (2006). Biomarkers from Huronian oil-bearing fluid inclusions: an uncontaminated record of life before the Great Oxidation Event. *Geology* **34**: 437–440.

Elvert M, Suess E, Greinert J, Whiticar MJ. (2000). Archaea mediating anaerobic methane oxidation in deep-sea sediments at cold seeps of the eastern Aleutian subduction zone. *Org Geochem* **31**: 1175–1187.

Feil C, Sussmuth R, Jung G, Poralla K. (1996). Site-directed mutagenesis of putative active-site residues in squalene-hopene cyclase. *Eur J Biochem* **242**: 51–55.

Fischer WW, Pearson A. (2007). Hypotheses for the origin and early evolution of triterpenoid cyclases. *Geobiology* **5**: 19–34.

Fischer WW, Summons RE, Pearson A. (2005). Targeted genomic detection of biosynthetic pathways: Anaerobic production of hopanoid biomarkers by a common sedimentary microbe. *Geobiology* **3**: 33–40.

Forster HJ, Biemann K, Haigh WG, Tattrie NH, Colvin JR. (1973). Structure of novel C35 pentacyclic terpenes from *Acetobacter xylinium*. *Biochem J* **135**: 133–143.

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hartner T, Straub KL, Kannenberg E. (2005). Occurrence of hopanoid lipids in anaerobic Geobacter species. *Fems Microbiol Lett* **243**: 59–64.

Hinrichs KU. (2001). A molecular recorder of methane hydrate destabilization. *Geochem Geophys Geosyst* **2**: art. no.-2000GC000118.

Hinrichs KU, Hayes JM, Sylva SP, Brewer PG, DeLong EF. (1999). Methane-consuming archaebacteria in marine sediments. *Nature* **398**: 802–805.

Hoshino T, Sato T. (2002). Squalene-hopene cyclase: catalytic mechanism and substrate recognition. *Chem Commun*, **4**: 291–301.

Hugenholtz P, Pace NR. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol* **14**: 190–197.

Mahenthiralingam E, Baldwin A, Drevinek P, Vanlaere E, Vandamme P, LiPuma JJ *et al.* (2006). Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* **1**: e17.

Mallory FB, Conner RL, Landrey JR, Zander JM, Greig JB, Caspi E. (1968). Biosynthesis of tetrahymanol from (4R)-[4-3H-2-14C]mevalonic acid. *J Am Chem Soc* **90**: 3564.

Ourisson G, Rohmer M, Poralla K. (1987). Prokaryotic hopanoids and other polyterpenoid sterol surrogates. *Annu Rev Microbiol* **41**: 301–333.

Pancost RD, Damste JSS, de Lint S, van der Maarel M, Gottschal JC, Medinaut Shipboard Sci P. (2000). Biomarker evidence for widespread anaerobic methane oxidation in Mediterranean sediments by a consortium of methanogenic archaea and bacteria. *Appl Environ Microbiol* **66**: 1126–1132.

Pearson A, Budin M, Brocks JJ. (2003). Phylogenetic biochemical evidence for sterol synthesis in the bacterium Gemmata obsuriglobus. *Proc Natl Acad Sci USA* **100**: 15352–15357.

Pearson A, Page SRF, Jorgenson TL, Fischer WW, Higgins MB. (2007). Novel hopanoid cyclases from the environment. *Environ Microbiol* **9**: 2175–2188.

Pearson A, Leavitt WD, Saenz JP, Summons RE, Tam CM, Close HG (2008). Diversity of hopanoids and squalene-hopene cyclases across a tropical land-sea gradient. *Environ Microbiol* (in press). doi:10.1111/j.1462-2920.2008.01817.x

Perzl M, Muller P, Poralla K, Kannenberg EL. (1997). Squalene-hopene cyclase from *Bradyrhizobium japonicum*: cloning, expression, sequence analysis and comparison to other triterpenoid cyclases. *Microbiology (UK)* **143**: 1235–1242.

Rappe MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630–633.

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008). Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**:1101–1104.

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.

Rohmer M, Bouviernave P, Ourisson G. (1984). Distribution of hopanoid triterpenes in prokaryotes. *J Gen Microbiol* **130**: 1137–1150.

Rohmer M, Ourisson G. (1976). Structure of bacteriohopanetetrols from *Acetobacter xylinium*. *Tetrahedron Lett*, **40**: 3633–3636.

Rusch DB, Halpern AL, 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.

Schulz-Gasch T, Stahl M. (2003). Mechanistic insights into oxidosqualene cyclizations through homology modeling. *J Comput Chem* **24**: 741–753.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stanier RY, Palleron NJ, Doudorof M. (1966). Aerobic pseudomonads—a taxonomic study. *J Gen Microbiol* **43**: 159–271.

Stepanauskas R, Sieracki ME. (2007). Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* **104**: 9052–9057.

Sugden MA, Talbot HM, Farrimond P. (2005). Flash pyrolysis—a rapid method for screening bacterial species for the presence of bacteriohopanepolyols. *Org Geochem* **36**: 975–979.

Summons RE, Bradley AS, Jahnke LL, Waldbauer JR. (2006). Steroids, triterpenoids and molecular oxygen. *Philos Trans R Soc B-Biol Sci* **361**: 951–968.

Summons RE, Jahnke LL, Hope JM, Logan GA. (1999). 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**: 554–557.

Talbot HM, Rohmer M, Farrimond P. (2007a). Rapid structural elucidation of composite bacterial hopanoids by atmospheric pressure chemical ionisation liquid chromatography/ion trap mass spectrometry. *Rapid Commun Mass Spectrom* **21**: 880–892.

Talbot HM, Rohmer M, Farrimond P. (2007b). Structural characterisation of unsaturated bacterial hopanoids by atmospheric pressure chemical ionisation liquid chromatography/ion trap mass spectrometry. *Rapid Commun Mass Spectrom* **21**: 1613–1622.

Talbot HM, Squier AH, Keely BJ, Farrimond P. (2003a). Atmospheric pressure chemical ionisation reversed-phase liquid chromatography/ion trap mass spectrometry of intact bacteriohopanepolyols. *Rapid Commun Mass Spectrom* **17**: 728–737.

Talbot HM, Summons R, Jahnke L, Farrimond P. (2003b). Characteristic fragmentation of bacteriohopanepolyols during atmospheric pressure chemical ionisation liquid chromatography/ion trap mass spectrometry. *Rapid Commun Mass Spectrom* **17**: 2788–2796.

Talbot HM, Summons RE, Jahnke LL, Cockell CS, Rohmer M, Farrimond P. (2008). Cyanobacterial bacteriohopanepolyol signatures from cultures and natural environmental settings. *Org Geochem* **39**: 232–263.

Talbot HM, Watson DF, Murrell JC, Carter JF, Farrimond P. (2001). Analysis of intact bacteriohopanepolyols from methanotrophic bacteria by reversed-phase high-performance liquid chromatography-atmospheric pressure chemical ionisation mass spectrometry. *J Chromatogr A* **921**: 175–185.

Talbot HM, Watson DF, Pearson EJ, Farrimond P. (2003c). Diverse biohopanoid compositions of non-marine sediments. *Org Geochem* **34**: 1353–1371.

Thiel V, Heim C, Arp G, Hahmann U, Sjovall P, Lausmaa J. (2007). Biomarkers at the microscopic range: ToF-SIMS molecular imaging of Archaea-derived lipids in a microbial mat. *Geobiology* **5**: 413–421.

Thiel V, Peckmann J, Richnow HH, Luth U, Reitner J, Michaelis W. (2001). Molecular signals for anaerobic methane oxidation in Black Sea seep carbonates and a microbial mat. *Mar Chem* **73**: 97–112.

Thoma R, Schulz-Gasch T, D'Arcy B, Benz J, Aebi J, Dehmlow H *et al.* (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432**: 118–122.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876–4882.

Tippelt A, Jahnke L, Poralla K. (1998). Squalene-hopene cyclase from *Methylococcus capsulatus* (Bath): a bacterium producing hopanoids and steroids. *Biochim Biophys Acta Lipids Lipid Metab* **1391**: 223–232.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wakeham SG, Amann R, Freeman KH, Hopmans EC, Jorgensen BB, Putnam IF *et al.* (2007). Microbial ecology of the stratified water column of the Black Sea as revealed by a comprehensive biomarker study. *Org Geochem* **38**: 2070–2097.

Wendt KU, Feil C, Lenhart A, Poralla K, Schulz GE. (1997). Crystallization and preliminary x-ray crystallographic analysis of squalene-hopene cyclase from *Alicyclobacillus acidocaldarius*. *Protein Sci* **6**: 722–724.

Wendt KU, Schulz GE, Corey EJ, Liu DR. (2000). Enzyme mechanisms for polycyclic triterpene formation. *Angew Chem Int Ed* **39**: 2812–2833.

Woodward RB, Bloch K. (1953). The cyclization of squalene in cholesterol synthesis. *J Am Chem Soc* **75**: 2023–2024.

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.

Zehr JP, Bench SR, Mondragon EA, McCarren J, DeLong EF. (2007). Low genomic diversity in tropical oceanic N-2-fixing cyanobacteria. *Proc Natl Acad Sci USA* **104**: 17807–17812.

Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward GF *et al.* (2001). Unicellular cyanobacteria fix N-2 in the subtropical North Pacific Ocean. *Nature* **412**: 635–638.

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW *et al.* (2006). Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**: 680–686.

Zundel M, Rohmer M. (1985a). Prokaryotic triterpenoids. 1. 3-beta-methylhopanoids from *Acetobacter* species and *Methylococcus capsulatus*. *Eur J Biochem* **150**: 23–27.

Zundel M, Rohmer M. (1985b). Prokaryotic triterpenoids. 3. The biosynthesis of 2-beta-methylhopanoids and 3-beta-methylhopanoids of *Methylobacterium organophilum* and *Acetobacter pasteurianus* spp. pasteurianus. *Eur J Biochem* **150**: 35–39.

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