Environmental Microbiology (2007) **9**(9), 2175–2188 doi:10.1111/j.1462-2920.2007.01331.x

Novel hopanoid cyclases from the environment

Ann Pearson,* Sarah R. Flood Page,[†] Tyler L. Jorgenson,[‡] Woodward W. Fischer and Meytal B. Higgins
Department of Earth and Planetary Sciences, 20 Oxford St., Harvard University, Cambridge, MA 02138, USA.

Summary

Hopanoids are ubiquitous isoprenoid lipids found in modern biota, in recent sediments and in low-maturity sedimentary rocks. Because these lipids primarily are derived from bacteria, they are used as proxies to help decipher geobiological communities. To date, much of the information about sources of hopanoids has come from surveys of culture collections, an approach that does not address the vast fraction of prokaryotic communities that remains uncharacterized. Here we investigated the phylogeny of hopanoid producers using culture-independent methods. We obtained 79 new sequences of squalene-hopene cyclase genes (*sqhC*) from marine and lacustrine bacterioplankton and analysed them along with all 31 *sqhC* fragments available from existing metagenomics libraries. The environmental *sqhC*s average only 60% translated amino acid identity to their closest relatives in public databases. The data imply that the sources of these important geologic biomarkers remain largely unknown. In particular, genes affiliated with known cyanobacterial sequences were not detected in the contemporary environments analysed here, yet the geologic record contains abundant hopanoids apparently of cyanobacterial origin. The data also suggest that hopanoid biosynthesis is uncommon: < 10% of bacterial species may be capable of producing hopanoids. A better understanding of the contemporary distribution of hopanoid biosynthesis may reveal fundamental insight about the function of these compounds, the organisms in which they are found, and the environmental signals preserved in the sedimentary record.

Received 9 January, 2007; accepted 2 April, 2007. *For correspondence. E-mail pearson@eps.harvard.edu; Tel. (+1) 617 384 8392; Fax (+1) 617 496 4387. Present address: [†]INSTAAR, University of Colorado, Boulder, CO 80309, USA; [‡]Arizona State University, Tempe, AZ 85287, USA.

Introduction

Molecular fossils are used by geologists and paleobiologists to reconstruct the history of Earth surface environments. The microbial lipids called hopanoids are among the most abundant of these molecules (Ourisson and Albrecht, 1992). Hopanoids and their degradation products are found in modern sediments (Farrimond *et al*., 2000), in oils of all geologic ages (Peters *et al*., 2005), and in Archean and Proterozoic rocks (Brocks *et al*., 1999; 2005; Summons *et al*., 1999; Xie *et al*., 2005). Their presence in modern and ancient sedimentary organic matter reflects resistance of the polycylic hydrocarbon skeleton to both biotic and abiotic degradation, making this class of lipids excellent biomarkers. In contemporary open-ocean and freshwater sediments, hopanoids are preserved in abundance. Export data from the Atlantic Ocean near Bermuda show that the water-column flux of hopanoids reflects episodic production in the local, overlying euphotic zone (Conte *et al*., 1998; 2003). Hopanoids preserved in ancient sedimentary rocks similarly are believed to reflect local depositional conditions and local microbial communities. Yet despite the importance of these compounds, the environmental, phylogenetic and/or physiological signals recorded by hopanoids remain poorly understood.

To the extent that preserved hopanoids can be linked to distinct sources, they can be used to interpret past microbial communities. For many years, the geohopanoids (hopanes) were regarded as 'orphan' biomarkers, because their origin was not known. In the 1970s, hopanepolyols – the primary biological precursors to geohopanoids – were found in *Acetobacter xylinium* (Förster *et al*., 1973; Rohmer and Ourisson, 1976). It was then proposed that hopanepolyols are functional analogues to eukaryotic sterols (Rohmer *et al*., 1979). The polycyclic structures, amphiphilic properties and molecular sizes of hopanepolyols might confer properties similar to sterols for regulating membrane rigidity and permeability (Ourisson *et al*., 1987).

To date, taxonomic information has come from surveys of species in culture (Rohmer *et al*., 1984; Summons *et al*., 1999; Talbot *et al*., 2001; Sugden *et al*., 2005). Rohmer and colleagues (1984) suggested that nearly 50% of prokaryotes contained hopanoids. Such a relatively common occurrence would be consistent with the hypothesized structural role for these compounds. The results of Rohmer and colleagues (1984) also suggested

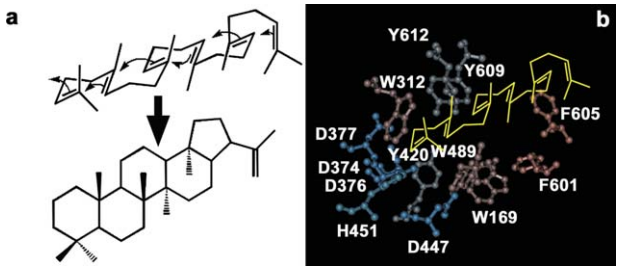

Fig. 1. A. Mechanism of the cyclization of squalene to form hopanoids.
B. The approximate location of pre-folded squalene relative to the critical AA loci (Hoshino and Sato, 2002) of the active site of SHC (*Alicyclobacillus acidocaldarius*; Wendt *et al*., 1997) as visualized using Cn3D-4.1 (http://130.14.29.110/Structure/CN3D/cn3d.shtml).

that hopanoids were restricted to obligate and facultative aerobes, although this now is known to be false. Hopanoid production occurs in anaerobic sediments and in common environmental groups such as the *Planctomycetales*, *Geobacteraceae* and *Desulfovibrios* (e.g. Pancost *et al*., 2000; Thiel *et al*., 2001; Sinninghe Damsté *et al*., 2004; Härtner *et al*., 2005; Fischer *et al*., 2005; Blumenberg *et al*., 2006).

In some cases, specific hopanoid structures do appear linked to phylogeny. Examples include an apparent association of 2-methylbacteriohopanepolyol with cyanobacteria, and hence a biomarker proxy for oxygenic photosynthesis (Summons *et al*., 1999). However, it remains unknown to what extent cultured species adequately represent the total diversity of hopanoid producers, as most of the global microbiota remain uncharacterized (Hugenholtz and Pace, 1996). A whole-community approach is needed if we are to understand the communities recorded by preserved geohopanoids.

To answer these questions, the environmental distribution of hopanoid synthesis genes can be investigated using molecular phylogenetic approaches. The squalene-hopene cyclase (*sqhC*) gene encodes the enzyme (SHC) that catalyses cyclization of hopanoids from their acyclic precursor, squalene (Fig. 1A; Ochs *et al*., 1992; Wendt *et al*., 1997). All SHCs must conserve the catalytic site for protonation of squalene. They must also control the stereochemistry of the pre-folded substrate and propagate the cyclization reaction. Several amino acid (AA) sequence motifs are conserved among SHCs of nearly all bacterial phyla to fulfil these strict requirements (Fig. 1B; Perzl *et al*., 1997; Hoshino and Sato, 2002; Fischer and Pearson, 2007).

In this study, we investigated *sqhC* diversity in three environmental samples. We used the conserved motifs to create third-codon degenerate primers to amplify partial *sqhC* genes using the polymerase chain reaction (PCR). *SqhC* represents a rare opportunity among biosynthetic genes: a limited set of primers will detect most of an entire phylogenetic Domain. The *sqhC* primer set is nearly 'universal': it is applicable to all characterized major groups of Bacteria, except for the divergent Planctomycetes. The data reported here more than double the total known number and diversity of *sqhC*s. Most of the new environmental sequences appear to represent novel taxa whose *sqhC*s are not currently represented in genomic databases. To our knowledge, this represents the first study of the domain-wide genetic diversity of a lipid biosynthesis gene in natural communities.

Results

PCR sequences

Putative *sqhC*s were sequenced from two aerobic photic zones and from an oxygen-depleted upper chemocline (~0.5–1 mg l^{-1} O_2). The samples were from the small, freshwater Lake Mishwam (LM) in Woburn, Massachusetts, USA; and from the oligotrophic North Central Pacific (NCP) Ocean. Microbial processes (e.g. Wick *et al*., 2000) have been studied extensively in LM; and the high organic load and stratified water column make LM similar to other small lakes that have been the subject of organic geochemical surveys, including for hopanoids (e.g. Priest Pot, UK; Innes *et al.*, 1997). The NCP sample was obtained as a sample of opportunity, in association with the work of Ingalls and colleagues (2006). Together they serve as representative fresh versus marine samples for this pilot study of *sqhC* diversity.

The data from LM contain 58 total clones representing 23 operational taxonomic units (OTUs) (Fig. 2A; Table S1). The NCP sample yielded only 21 clones representing 11 OTUs (Fig. 2C; Table S1). The sequences reflect the application of up to eight different combinations of primer pairs and PCR amplification programmes (Fig. S1). Many of the sequences that occurred multiple times were found using more than one of the combinations. This suggests that repeat occurrences of the same clone are not solely the result of amplification biases contributed by one or more of the primer sets or methods. The most frequently detected sequence, clone *s* from LM, appeared 21 times from five methods with the following distribution: 9, 4, 3, 3, 2. Despite the many different trials, most data derive from just three primer sets: a single forward primer, with either cyR, pr_R or prAR. Reverse primer prBR is only one nucleotide different from cyR and the results of this primer overlap with results from cyR (Table S1).

The diversity of the data was estimated from the DNA sequences using the DOTUR statistical package (Schloss and Handelsman, 2005). Rarefaction curves are shown for LM (Fig. 2B) and NCP (Fig. 2D). DOTUR allows rarefaction analyses using multiple different thresholds of similarity cut-off to define OTUs. The data are shown for 6% and

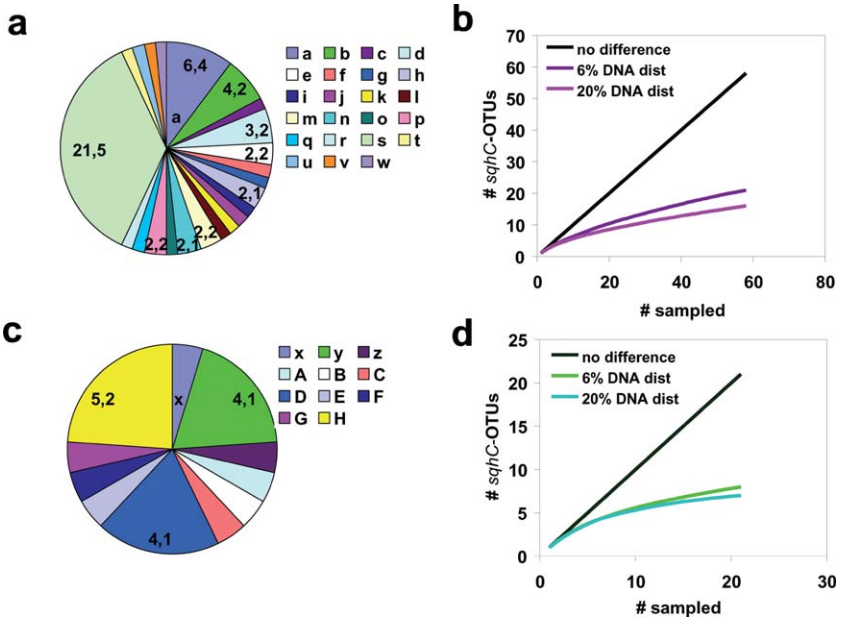

Fig. 2. A and C. Pie charts of *sqhC* clones sequenced from Lake Mishwam (A) and the North Central Pacific (C). Numbers (x,y) show clone frequency (x) and in how many (y of 8) methods the clone appeared; segments without numbers are by default (1,1). Clones appear alphabetically in clockwise order; letters refer to Table S1.
B and D. Rarefaction curves for *sqhC* DNA sequence data from Lake Mishwam (B) and the North Central Pacific (D) calculated using DOTUR as described in *Experimental procedures*.

20% cut-offs and for the theoretical 0% (all sequences unique). These are the same thresholds as used for the Sargasso Sea *rpoB* gene data when these data were analysed using DOTUR (Schloss and Handelsman, 2005). By these methods, *rpoB* genes have OTU diversity approaching 10^3 in the Sargasso Sea, while *sqhC* genes in the NCP and in LM have OTU diversity $< 10^2$.

Metagenomes

All putative *sqhC*s from environmental metagenomics databases are listed in Table S2. The original Sargasso Sea (SS) data set (Venter *et al*., 2004), which contains ~1 Gb of unique sequence (Table 1), contains 25 genes or fragments that when translated into AA sequence are homologous either to SHCs or to sterol oxidosqualene cyclases (OSCs). The mean exponent of the expectation value for these sequences is −72 (i.e. $E = 10^{-72}$) for TBLASTN. The AA sequence identity for environmental SHCs and their nearest relatives among genomic sequences ranges from 38% to 96% (mean 59%). Squalene-hopene cyclases and OSCs share lower similarity (e.g. the SHC and OSC of *Methylococcus capsulatus* BATH are 23% identical to each other).

In comparison with the SS community, the data available from the Diversa silage farm soil (FS; Tringe *et al.*, 2005) metagenome are more limited. Approximately 10^5 genes (total of 152 Mb of DNA; Table 1) have been depos-

Table 1. Summary statistics for metagenomes.

	No. *sqhC*	Total size (Mb)	Genomes	*sqhC* per genome	OTUs
Sargasso Sea	23[a]	1045	522	0.04	> 1800
Farm soil: Diversa silage	5	152	76	0.07	> 3000
Acid mine drainage	3[b]	11	5	0.60	6
Whale fall sample #1	0	32	16	< 0.06	
Whale fall sample #2	0	32	16	< 0.06	
Whale fall sample #3	0	31	16	< 0.06	
Sludge/USA	0	28	14	< 0.07	
Sludge/Australia	0	27	14	< 0.07	
Human gut #8	0	20	10	< 0.10	
Human gut #7	0	16	8	< 0.10	
Monterey Bay Coastal Observatory	0	5	2.5	< 0.40	

a. Twenty-five putative triperpenoid cyclases were detected, of which 23 are putative SHCs and 2 are putative OSCs.
b. There are two copies of *sqhC* in *Lepto*. gp. II; one copy in *Lepto*. gp. III. Shotgun data were binned into five groups, three from archaea plus the two *Leptospirillum* spp. *Sulfobacillus* (1% of population) data are folded into *Lepto*. gp. III; thus the number of OTUs is six, despite only five complete or partial genomes sequenced (Tyson *et al*., 2004).
Columns: (2) number of *sqhC* fragments of length 225 bp or longer (translated length 75 AA); (3) total unique sequence in 10^6 bp (Mb); (4) full genome equivalents, assuming 2 Mb per genome; (5) frequency of *sqhC* per environmental genome equivalent of total sequence; (6) estimated total community 16S rRNA taxonomic diversity (OTUs).

ited in the public databases. Five *sqhC* fragments were detected within these data. In the acid mine drainage site from Iron Mountain, CA (IM; Tyson *et al.*, 2004), putative *sqhC*s were found in the genomes of both *Leptospirillum* group II and *Leptospirillum* group III (*Leptospirillum ferrodiazotrophum*). The TIGR-Monterey Bay Coastal Ocean Microbial Observatory sequencing project (http://www.tigr.org/tdb/MBMO/MBMO.shtml) contains no apparent *sqhC*s. Similarly, no *sqhC* homologues were found in the human gut communities (Gill *et al*., 2006), whale falls samples (Tringe *et al*., 2005) or sewage sludge samples (Martin *et al*., 2006).

The proportional abundance of *sqhC*-containing organisms in these environmental communities appears to be small. The Sargasso Sea *sqhC*s represent five sequences on assembled scaffolds and 18 paired-end reads (single coverage) or singletons. Of the scaffolded sequences, three are from the unexpected *Burkholderia* spp. in the SS that may reflect contamination (DeLong, 2005; Mahenthiralingam *et al*., 2006). The remaining two scaffolded sequences represent organisms abundant enough to have their genomic DNA appear more than once in the total 1625 Mb of raw sequence. However, most *sqhC* genes in the SS sample occur a single time. Thus, the number of unique *sqhC* genes is approximately equal to their total incidence of detection. Ribosomal RNA genes occur in this data set ~1400 times, and the total species richness (shown in the OTU column, Table 1) was estimated by various methods to be > 1800. The ratio of *sqhC* genes to either of these numbers yields the prediction that ~2% of species in the SS would contain *sqhC* (assuming one copy per genome). Alternatively, using an estimated average genome size of 2 Mb, the 1045 Mb of total unique sequence equals 522 genome length equivalents. Using this number, instead, would predict that *sqhC* occurs in ~4% of SS genomes.

The most abundant taxon in the total metagenome of the Diversa farm soil is covered at < 1X (Tringe *et al*., 2005). All of the FS *sqhC* fragments are unique. If 152 Mb of sequence is equivalent to 76 total genomes in length, *sqhC* genes occur in ~7% of the genomes in this surface soil, although this must remain a tentative estimate (Table 1). In all of the other metagenomes, only the upper limit can be estimated for the frequency of *sqhC*. This number is determined primarily by the length of sequence deposited, but again it is consistent with the presence of *sqhC* in < ~10% of both total genome length equivalents and total unique genomes, except for the unusual acid mine drainage community.

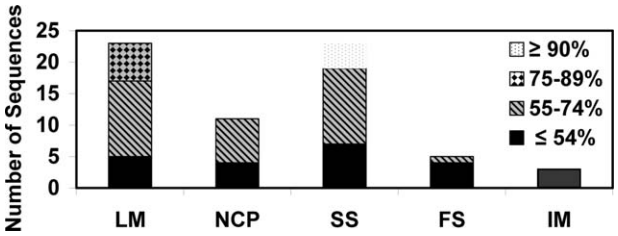

Fig. 3. Per cent identity of the environmental SHC AA sequences relative to their respective best matches in GenBank. All data from clone libraries (LM, NCP) and metagenomes (SS, FS, IM) are shown. The four sequences from the SS metagenome having ≥ 90% identity all belong to the *Burkholderia* spp. that are suspected contaminants of the SS sample.

Similarity to known SHCs

All of the translated SHCs are novel: none can be assigned to a specific species and the vast majority not even to a known genus. This is true both for the PCR amplicons and for the metagenomics fragments. Despite the broad diversity of > 60 SHC reference sequences representing > 37 genera currently in GenBank, the LM clones average 62% AA identity with their best matches in the genomic database, and the NCP clones average 59% identity (Table S1). The SHCs from metagenomes are similarly unique: the AA sequences of the fragments, most of which are from the Sargasso Sea and are comparable to the NCP, also average 59% identity with their closest known relatives (Table S2). Among the environmental data, clone ***I*** from LM has the closest match to a known organism: it shares 86% AA identity with the genus *Pelobacter*.

In comparison, SHCs generally are ≥ 90% identical within a genus (e.g. *Nitrobacter winogradskyi* versus *N. hamburgensis*); ≥ 75% identical within a subgroup classification such as α-proteobacteria (e.g. *Nitrobacter* versus *Rhodopseudomonas*); 50–60% identical between subgroups of a major group (e.g. α-proteobacteria versus γ-proteobacteria); and only 40–55% identical between unrelated major groups (e.g. α-proteobacteria versus cyanobacteria). Using these approximate guidelines, most of the environmental SHCs would belong to taxa that are not yet represented within the genomic databases and/or in pure culture. Nearly 40% of the SHCs fall below the 55% similarity threshold and would not be classified as belonging to any major bacterial group yet described (Fig. 3). Another 50% of the sequences are 55–74% similar to a known organism and may be classifiable within a major group (e.g. proteobacteria or Actinomycetes) but would not place within a subdivision such as α-proteobacteria. Only 10% of the sequences exceed the 75% similarity threshold. All are from LM and fall within the δ-proteobacteria (Table S1).

Phylogenetic trees

Two trees were created to examine the SHC sequences. The first uses full-length SHCs (~620 AA) from species in

pure or enrichment culture (Fig. 4A). Nine representative sterol OSCs were included to form a natural outgroup cluster (Pearson *et al*., 2003; Summons *et al*., 2006). The sequences formed 12 major taxonomic groups: (i) α-proteobacteria type I, (ii) primarily β-proteobacteria, (iii) α-proteobacteria type II, (iv) β- and γ-proteobacteria, (v) δ-proteobacteria type A, (vi), δ-proteobacteria type B, (vii) δ-proteobacteria type C, (viii) Gram-positive bacteria (Actinomycetes), (ix) cyanobacteria, (x) the anammox planctomycete, *Candidatus* 'Kuenenia stuttgartiensis', (xi) Acidobacteria and (xii) other Planctomycetes. Only two nodes in this tree do not receive at least 50% bootstrap support using three different AA substitution matrices for likelihood analysis.

The second tree includes the shorter environmental sequences (PCR and metagenomes; Fig. 4B). The portion of the enzyme that falls between the translated AA motifs of the forward and reverse PCR primers was used. This minimized the effects of length heterogeneity among the metagenome sequences. Because of this restriction, only some of the metagenome sequences were incorporated (those that overlapped this region). Loss of bootstrap support and notable rearrangements of the taxonomic clusters was observed despite this conservative approach, possibly caused by the high sequence dissimilarity between the environmental samples and the whole-genome data (Tables S1 and S2). Several new taxonomic clusters appear in Fig. 4B which are predominantly or entirely composed of novel SHCs.

Groups 1–4 are re-ordered in Fig. 4B with respect to Fig. 4A. Environmental sequences within these groups include the *Burkholderias* from SS. Three additional sequences, two from SS and one from NCP, are basal to groups 1–4. While these SHCs are most closely related to *Rhodospirillum* by TBLASTN (Altschul *et al*., 1997), they have only 59–66% AA identity with this genus and do not place in group 3. Two sequences from NCP may be related to δ-proteobacteria (group 5), but are < 70% identical to *Geobacter* and *Pelobacter*. The two from the acid mine drainage (1Fe_LII, 2Fe_LIII) belong to *Leptospirillum* spp., phylum *Nitrospirales*, but bootstrap support is poor for the other environmental sequence (1FS) basal to group 5. Its phylogenetic affiliation remains unknown. A large cluster from LM is associated with group 6 (δ-proteobacteria). This cluster includes clone ***I***, the sequence related to *Pelobacter*.

Elsewhere, the phylogeny of the environmental SHCs is uncertain. A cluster from NCP falls between groups 6 and 7, but is 58–69% identical (Table S1) to known sequences from α-proteobacteria. This illustrates the need for caution when making phylogenetic assignments based on similarity scores and affiliations as reported by BLAST. Group 12 contains Planctomycetes and includes one sequence from SS and two from FS. The only remaining environmental SHC that appears related to a known organism at or above the major group level is clone TJ2cySn10, which groups with the Acidobacteria. All other SHCs that fall between groups 9 and 11 form clusters distinct from any previously characterized phylum. There also are no environmental sequences that appear to be affiliated with the cyanobacteria, either phylogenetically or by having AA identity > 55%. The closest candidate is LM clone TJ2cySn16, which is 50% identical to *Gloeobacter* but clusters between groups 11 and 12.

Discussion

Distribution of hopanoids in cultivated microbes

While all Eukaryotes require either biosynthetic or dietary sterols, a seminal study of the distribution of hopanoids showed they were not ubiquitous among bacteria, but rather were present in ~50% of the > 90 strains surveyed (Rohmer *et al*., 1984). Currently, of the > 600 Bacteria that have had their genomes sequenced, ~10% contain *sqhC* genes (Fischer and Pearson, 2007). This suggests a lower frequency of hopanoid biosynthesis among Bacteria than originally reported. However, an important concern is the extent to which cultured organisms represent the distribution of biochemical pathways in nature, as it is commonly accepted that < 1% of environmental species can be grown in the laboratory. Both of the above estimates of the frequency of hopanoid biosynthesis among bacteria (50% or 10%) are equally uncertain. Investigation of *sqhC* genes in nature is required, as it is the integrated environmental population that is responsible for the biomarker record in sediments.

Early taxonomic information also showed few phylogenetic or metabolic affinities among hopanoid-producing organisms (Rohmer *et al*., 1984). One significant association appeared to be 3β-methylhopanoids, which derive from microaerophilic methanotrophs, methylotrophs and acetic acid bacteria (Zundel and Rohmer, 1985; Summons and Jahnke, 1992). They often are depleted strongly in ^{13}C (Jahnke *et al*., 1999), confirming that methanotrophs are important sources of these lipids. In another important example, the 2-methylhopanoids are believed to be good fossil markers for cyanobacteria (Brocks *et al*., 1999; Summons *et al*., 1999; Xie *et al*., 2005). In general, however, there remain many questions about the non-universal distribution of hopanoids. How representative is this prior state of knowledge?

Diversity and frequency of environmental sqhCs

Both the PCR and metagenomic data suggest that the biosynthesis of hopanoids is a rare pathway among

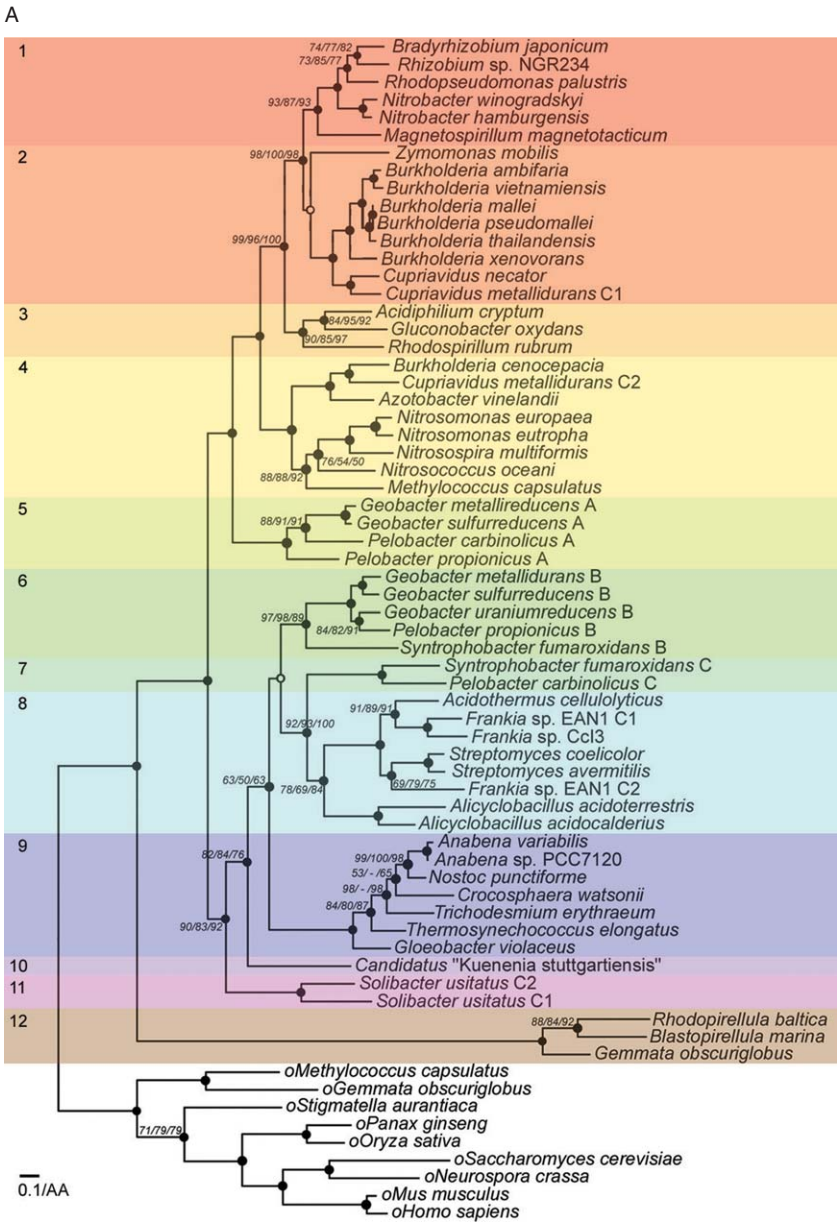

Fig. 4. Maximum likelihood trees of SHC and OSC sequences from characterized bacterial species, as well as selected OSC sequences of eukaryotes (e.g., *Homo sapiens*) (A) and from alignable environmental sequences (B). Bootstrap values are shown for JTT/Dayhoff/WAG substitution matrices in (A) and JTT only in (B). Nodes marked with solid circles indicate > 50% bootstrap support in all calculations; no number next to the node implies 100/100/100 support. Nodes marked with empty circles indicate < 50% bootstrap support in one or more calculation. In (A), the tree topology is identical in all cases except Dayhoff matrix calculation, in which *T. erythraeum* and *C. watsonii* are siblings. In (B), rearrangement occurred as noted by comparing the order of colour-coded groups. Environmental sequences: blue (Sargasso Sea); red (Acid Mine); brown (Farm Silage); green (North Central Pacific); purple (Lake Mishwam). White gaps in (B) correspond to environmental SHC sequences with no known affiliates among characterized species. Single-letter suffixes for LM and NCP sequences refer to the clone letters in Fig. 2. Accession numbers are shown for all SS, FS and IM sequences.

the bacteria. Despite the widespread preservation of hopanoids in the geologic record, the metagenome data show that hopanoid cyclase genes are sparsely distributed in the environments examined to date. The true fraction of hopanoid producers in most communities may be no more than 10%, and in aquatic systems it could be < 5% (Table 1). In the total of 1.4 Gb of unique DNA sequence available from all metagenomics projects (as of November 2006), there are 31 complete or partial *sqhC* genes (Table S2). The SS metagenome includes an estimated 1.2 million genes and gene fragments. The 23 *sqhC*s in the SS metagenome contrast with the presence of 782 putative bacterial rhodopsins in the same data (Venter *et al*., 2004). Hopanoid-producing organisms could be as much as a factor of 30 less abundant than light-harvesting bacteria in the surface ocean. Such a rare occurrence may not be consistent with the fundamental membrane-structural role proposed for hopanoids (e.g. Ourisson *et al*., 1987), because so few bacteria apparently produce these compounds.

In contrast to the SS, nearly every bacterial cell living in the acid mine drainage (IM) community is expected to produce hopanoids. Iron Mountain is an unusual 'extreme' environment, so despite the possibility that SHCs could be common in acidic environments, it is unknown to what extent that site could be considered representative of other terrestrial systems. A more typical example may be the farm soil metagenomic

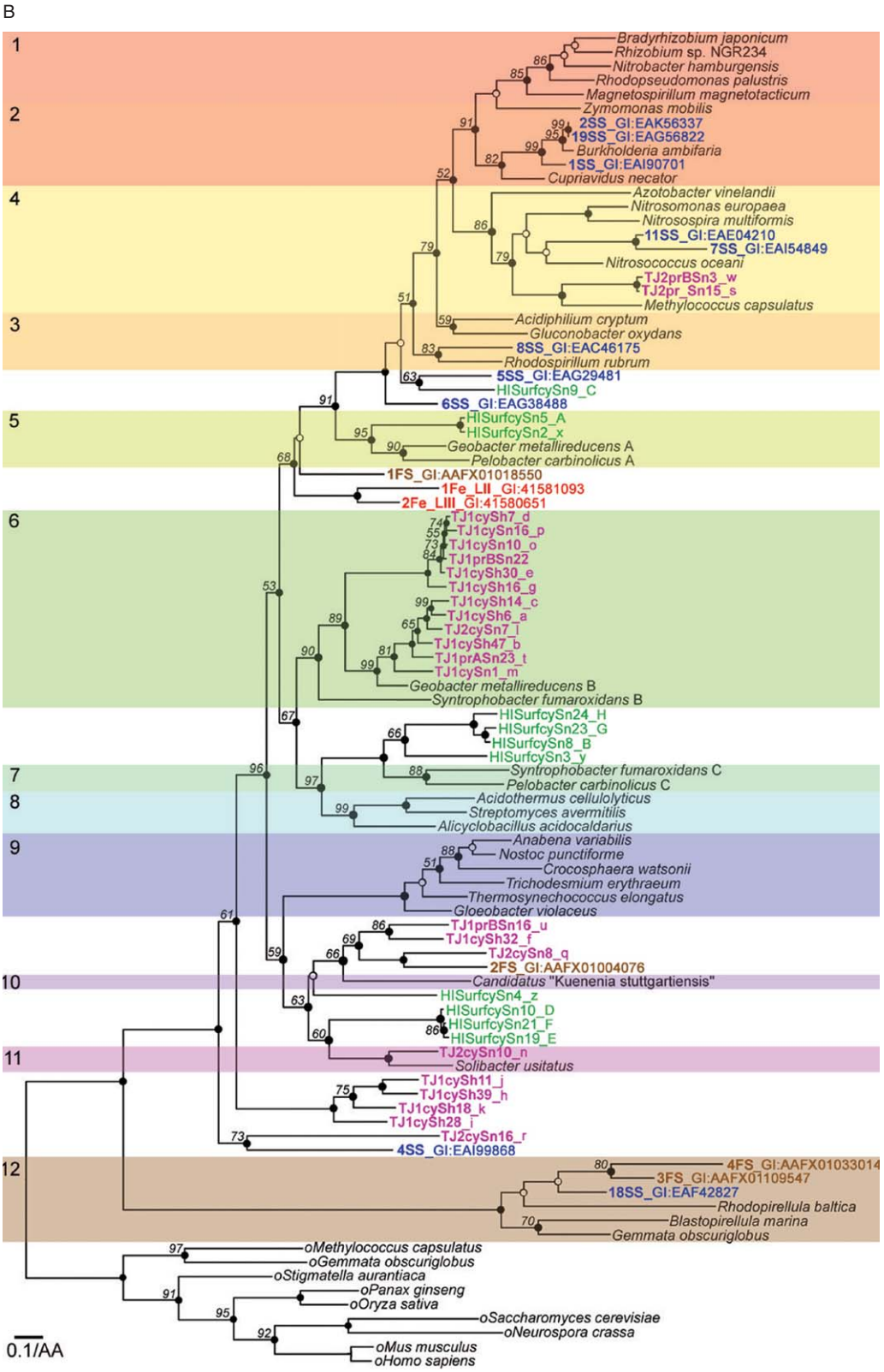

Fig. 4. *cont.*

community (Tringe *et al*., 2005). Although only 100 000 genes have been sequenced and deposited in the public databases, it is estimated that this community is more than twice as diverse as the Sargasso Sea. Table 1 suggests that the *sqhC* gene density also is greater in the FS sample than in the SS (present in 7% instead of 2–4% of genome equivalents). Tentatively, this may suggest that proportionally more of the total cells and/or species in terrestrial communities contain *sqhC* genes.

To date, shotgun and large-insert sequencing projects have yielded fewer than half the number of sequences as were obtained by direct amplification (31 versus 79). The diversity of sequence obtained by PCR also is at least as great as the diversity obtained by whole-genome shotgun sequencing. The sequences obtained from LM and NCP are as widely distributed phylogenetically (Fig. 4B), as are the metagenomics SHCs. However, a significant concern is that both types of data cover a narrow spectrum of environments. More work is needed across a representative range of terrestrial and aquatic systems.

Other analytical concerns include to what extent the PCR and metagome data may both be biased in their recovery of sequence diversity. The searches of environmental metagenomes should detect all SHC homologues in these databases, because the high degree of conservation permits comprehensive detection by similarity to known SHC sequences. Squalene-hopene cyclases have expectation values during similarity searches (e.g. TBLASTN; Altschul *et al*., 1997) usually $< 10^{-70}$ for full-length sequences (Perzl *et al*., 1997; Fischer and Pearson, 2007). The enzyme with the closest structural (Wendt *et al*., 1997; Thoma *et al*., 2004) and sequence (Pearson *et al*., 2003; Sawai *et al*., 2006; Summons *et al*., 2006) homology to SHC is sterol OSC. We detected two such fragments in the SS data set, which were identified as putative OSCs by differences in certain critical AA motifs (Table S2).

The presence or absence of key AA residues corresponds directly to the enzymatic function. The catalytic site of protonation in all SHCs is 374-DxDD-377 (Hoshino and Sato, 2002), and there are no known exceptions among genera sequenced to date (Fischer and Pearson, 2007). Additional criteria for inclusion of a putative SHC were aromatic residues F601 and F605 (Fig. 1B), if the fragment overlapped these loci upon alignment. Residues F601 and F605 are required to form the hopanoid cyclohexyl C and D rings respectively (Pale-Gosdemange *et al*., 1998; Merkofer *et al*., 1999; Hoshino *et al*., 2000). The presence of these motifs and other patterns (e.g. six structural repeats of the form QxxxGxW) is diagnostic for SHCs (Wendt *et al*., 1997; Hoshino and Sato, 2002). These selection criteria minimized the likelihood of retaining false positives.

In comparison, it is more likely that the new sequences obtained by PCR have not yielded the full diversity of SHCs from these environments. Amplification by PCR can cause proportional biases during amplification (e.g. Polz and Cavanaugh, 1998). The primers used here also did not target Planctomycetes; nor would any of the primer combinations capture taxa similar to sequence 9SS, which contains the AA motif SPVWDS in the forward primer location (rather than SPVWDT). The phylogenetic diversity of the environmental amplicons (Fig. 4B) does suggest, however, that the approach samples a wide breadth of the species containing *sqhC* genes. The total projected diversity (< 100 unique SHCs per environment) and overall degree of phylogenetic distance (typically ~60% identical to nearest relatives) also are similar between the PCR results and the metagenomic data, again suggesting the PCR data are not overwhelmed by amplification biases. However, further work will be needed to address these issues of comprehensiveness and quantitation more explicitly.

Environmental taxonomy

The distribution of *sqhC*-containing organisms in Fig. 4 confirms many of the original conclusions of Rohmer and colleagues (1984). *SqhC* universally is absent from the Archaea and from known species of purple and green sulfur bacteria. Among other Bacteria, there is heterogeneity within the major phylogenetic divisions; e.g. the marine cyanobacterium, *Trichodesmium erythraeum* contains *sqhC*, while *Prochlorococcus* spp. do not. Similarly, among the α-proteobacteria, *Rhodopesudomonas* spp. can produce hopanoids, while *Rhodobacter* spp. cannot. While only ~30% of the strains from the Rohmer and colleagues (1984) project have been sequenced to date, among them, the hopanoid-positive species contain *sqhC* genes, and the hopanoid-negative species do not. This suggests that possession of *sqhC* generally implies its expression. Known exceptions are the case of a *Rhizobium* sp. (NGR234; Freiberg *et al.*, 1997), in which *sqhC* is present on a plasmid (*sqhC*[+]), although *Rhizobium* spp. appear not to make hopanoids; *Bacillus* spp., which do not produce hopanoids, although they contain *sqhC* homologues (*sqhC*[+/–]; the [–] signifying critical mutations); and *Streptomyces coelicolor*, which expresses its *sqhC* only when forming aerial hyphae (Poralla *et al*., 2000).

Most of the sequences from organisms in culture are proteobacterial (Fig. 4A; groups 1–7). This may reflect bias in the types of organisms selected for pure culture and/or sequencing; it does not suggest that most hopanoid producers, globally, are necessarily Proteobacteria. The environmental clones placing within and between groups 1–7 (Fig. 4) may reflect affiliation with α-,

β-, γ- or δ-*proteobacteria*, but many of the individual sequences are too distant to be classified into these defined groups. It remains unknown if this reflects greater AA sequence dissimilarity among *proteobacteria* than outlined in the taxonomic guidelines above, or if these SHC sequences do not derive from *proteobacteria*. For example, the sequences from the phylum *Nitrospirales* cluster with SHCs from *proteobacteria*, falling between SHC groups 5 and 6. It is likely that many of the novel environmental SHCs belong to other, unidentified phyla. This is especially probable at the base of the tree, where the SHCs between groups 9 and 12 more often have lower percentage AA identities (Tables S1 and S2).

Some environmental SHCs do appear to have informative taxonomic associations, however. The most abundantly detected clone, ***s***, appears related to *Methylococcus* spp. (Fig. 4). Lake Mishwam is methanogenic. A sharp density contrast maintains permanently anaerobic conditions and stratification is preserved during all but the most extreme weather systems (Wick *et al*., 2000). Methane is oxidized at the chemocline boundary, and aerobic methanotrophs (γ- and α-proteobacteria) are found in the water column (M. Fisher, pers. comm.). Clones ***s*** and ***w*** probably belong to aerobic methanotrophs within the γ-proteobacteria. Methanotrophs produce hopanoids (Rohmer *et al*., 1984; Summons and Jahnke, 1992; Talbot *et al*., 2001) and contain *sqhC*s (Tippelt *et al*., 1998).

A large cluster of sequences, also from LM, groups with the δ-proteobacteria group 6. This includes clone ***l***, which is 86% identical to *Pelobacter*. If these clones from LM indeed represent other genera or species related to *Pelobacter*, such results are intriguing, as this genus is anaerobic. However, *Geobacter sulphurreducens* can grow under conditions of limited O_2 (Lin *et al*., 2004). Alternatively, the LM clones may derive from uncultivated members of the *Myxococcales* or *Bacteriovorax* lineages of δ-proteobacteria, both of which contain aerobes. It also is possible that these samples contain particulate material mixed in from deeper waters. Although the lake stratification is extreme (surface water residence time a few days, deep water residence time > 100 days; Wick *et al*., 2000) some cross-chemocline mixing must occur. Quantitative PCR and comparisons between surface and deep waters may resolve this question in the future.

Implications

One of the primary goals of geobiology is to reconstruct and understand the co-evolution of life and environment. When did important biogeochemical pathways such as oxygenic photosynthesis first leave an impact on the geologic record? In sedimentary rocks, some of the most ancient records of biological activity are obtained from lipid biomarkers (e.g. Brocks *et al*., 1999; Dutkiewicz *et al*., 2006). On any time-scale, the most valuable biomarkers are both taxonomically and environmentally diagnostic, i.e. they can be assigned to a defined group of organisms and record a specific environmental signal. Examples include the sea surface temperature proxies TEX_{86} and $U_{37}^{k'}$, which are based on ether lipids of *Archaea* (Schouten *et al*., 2002) and alkenones of haptophyte algae (Brassell *et al*., 1986; Prahl and Wakeham, 1987) respectively. Other specific lipids reveal the timing of major evolutionary radiations, such as the distribution of highly branched isoprenoids (HBI) of diatoms (Sinninghe Damsté *et al*., 2004).

While some applications of biomarker proxies derive from empirical correlations between lipids and depositional conditions (Peters *et al*., 2005), other geobiological applications of biomarkers necessarily proceed from knowledge of the organisms that produce them. Yet because the vast majority of prokaryotic diversity remains uncharacterized, this could represent a proportional reservoir of unknown biosynthetic diversity. The data suggest this is true for the producers of hopanoids. The novel sequences for putative squalene-hopene cyclases reported here (Tables S1 and S2; Fig. 4B) confirm the presence of numerous uncultured – and therefore unstudied – species that may be capable of hopanoid biosynthesis. The specific lipids produced by these environmental taxa remain unknown. For example, uncultured, novel phyla may produce 2-methyl- or 3-methylhopanoids. We lack phylogenetic and physiological information about these undiscovered organisms.

The samples examined here only represent aerobic surface environments from subtropical and temperate latitudes. The diversity of hopanoid producers in anaerobic water columns and sediments remains unexplored. Further work should include more surveys of environmental metagenomes from anaerobic sediments, euxinic water columns and terrestrial bogs, as well as a broader range of climatic and depositional regimes. The metagenomes also suggest there may be differences between terrestrial and aquatic systems. *SqhC*s appear twice as frequently per Mb of sequence in the farm soil metagenome (Tringe *et al*., 2005) as in the Sargasso Sea (Venter *et al*., 2004).

Little is known about the phenotypes conferred by hopanoids, but this information is needed to help explain differing patterns of hopanoid gene distribution, whether in association with oxygen concentrations or other environmental controls. Among bacteria that do produce hopanoids, some only do so during a specific life-cycle stage or for a specific ecological purpose. The nitrogen-fixing actinomycete *Frankia* forms hopanoid vesicles that are believed to protect nitrogenase from exposure to oxygen (Berry *et al*., 1993; Kleemann *et al*., 1994; Nalin *et al*., 2000). In general, nitrogen fixation is common but

not universal among bacteria with *sqhC* genes. For example the *Leptospirillum* groups II and III bacteria from IM both contain *sqhC* genes, but only *Leptospirillum* III fixes N_2 (Tyson *et al*., 2004). Under what conditions might hopanoids help protect organisms from oxygen stress? A role for triterpenoids in oxygen protection would have both evolutionary and geological implications. The non-uniform and rare distribution of hopanoid biosynthesis hints at a richer biological function of these compounds than currently is understood. If these roles can be deciphered, there may be new applications for hopanoids as environmental proxy biomarkers.

There are also known producers of hopanoids that are surprisingly absent from the current data. Cyanobacteria are believed to be major sources of the hopanoids found in sedimentary rocks during several geologic ages (Brocks *et al*., 1999; Summons *et al*., 1999; Xie *et al*., 2005). Specifically, degradation products of 2-methylbacteriohopanepolyols are attributed to cyanobacteria. High concentrations of these compounds are believed to reflect the dominance of prokaryotic oxygenic photosynthesis in the ancient geologic record. However, none of the five samples investigated here contain *sqhC*s that can be attributed to cyanobacteria. This suggests that hopanoids in such environments as the Sargasso Sea do not derive from cyanobacterial production, which is especially important considering the overwhelming evidence that the source of microbial biomass exported to sediments is local (e.g. Conte *et al*., 2003).

Such a result certainly prompts a major question: cyanobacteria that produce hopanoids seem to have been widespread in the geologic past, so where are they now? Some, but not all species of cyanobacteria contain *sqhC* genes (Fig. 4A). In which contemporary ecological settings and environments might we find hopanoid-producing cyanobacteria in abundance? Have hopanoid-producing cyanobacteria largely been succeeded, globally, by non-hopanoid producers (e.g. *Prochlorococcus* and *Synechococcus*)? If so, why? The answer may provide fundamental insight about the history of Earth over geologic time. The ancient sediments that contain abundant 2-methylhopanoids from cyanobacteria may record an Earth surface environment that fundamentally is different geochemically and/or climatically from modern conditions.

Experimental procedures

Environmental samples

Two samples of suspended particulate matter (SPM) from LM were obtained by gentle vacuum filtration through 0.2 μm cellulose nitrate filters (Millipore™). Filters were stored at –80°C until total DNA was extracted using the UltraClean™ Mega Soil DNA Kit from MoBio Labs. DNA was concentrated and cleaned using a Qiagen™ QIAquick purification kit. Two samples, TJ1 and TJ2, represent 1.5 l each of surface (1.8 m) and upper chemocline (2.4 m) waters respectively.

The sample from the North Central Pacific was collected at the Natural Energy Laboratory of Hawaii Authority (NELHA). Seawater from 21 m depth was collected onto 0.2 μm Pall Supor® filters and frozen at –70°C for return to the laboratory. The complete sampling description is given in Ingalls and colleagues (2006). All samples were lysed in 1.5 M Na-perchlorate at 4°C for 48 h (Blair *et al*., 1985). DNA was extracted using phenol:chloroform:isoamyl alcohol (25:24:1) by aliquoting the lysate into 50 ml Falcon® tubes. DNA was concentrated and cleaned as above.

Primer design

Primers were designed to bound the active catalytic site of protonation (PDxDD) conserved in all SHC enzymes. SP(V/I)WDT is a motif approximately 65 AA in the N-terminal direction, DGGWGE is a motif approximately 160 AA in the C-terminal direction, and T(G/A)TGFP, NA(V/P)GFP and (F/Y)FPL(W/Y)A are a series of motifs approximately 230 AA in the C-terminal direction from the catalytic site. We designed six degenerate primers based on these motifs (primer sequences and conservation among described bacterial genera is given in Tables S4 and S5).

PCR amplification

Partial *sqhC* genes were amplified from the three environmental DNA samples by using different combinations of primers and temperature programmes (Fig. S2, Fig. S3 and Table S3). Between five and eight different combinations were attempted per sample. Full clone names reflect the combination of sample–primer method; e.g. TJ2prBSn means sample TJ2, using reverse primer prB, with the combination touchdown-nested (Td-N) PCR method, n. All clones in Table S1 and Fig. 4B follow this notation. Some of the method combinations yielded no sequences, usually through failure to generate any SHC sequences among insert-containing clones (27% of inserts were correctly *sqhC*; the remainder were non-specific amplicons). Thus, Table S1 does not contain prefixes matching all possible combinations (e.g. there are no clones for HISprASn that appear in Table S1, although this combination was attempted). Details of PCR methods and thermocycler programmes appear in *Supplementary material*.

Cloning and sequencing

Polymerase chain reaction products from all samples were cloned using the Invitrogen TOPO TA Cloning® Kit for Sequencing with One Shot® TOP10 Chemically Competent *Escherichia coli*. Clones were incubated overnight at 37°C on LB Media/Ampicillin (50 μg ml^{-1}) plates. Colonies were inoculated into 5 ml of liquid LB Media/Ampicillin (50 μg ml^{-1}) and incubated for 24 h at 37°C. Plasmids were purified using the Qiagen QIAprep® Spin Miniprep Kit. DNA was quantified on a Beckman Coulter DU® 640 Spectrophotometer and when necessary, re-concentrated using ethanol precipitation. Sequencing was done at the Dana-Farber/Harvard Cancer Center DNA Resource Core (http://dnaseq.med.harvard.edu/).

Primer controls

Negative controls (reagents, lab water) were run with every PCR reaction, and contamination never was found. Positive genomic controls were run using DNA from *G. sulphurreducens* PCA, *Nostoc punctiforme* PCC73102, *Rhodopseudomas palustris* CGA009 and *Methylococcus capsulatus* BATH. Both 'cyanobacterial' (cy) and 'proteobacterial' (pr_, prA, prB) primer combinations worked to variable extents on multiples of these organisms, consistent with the similar sequences and non-specificity of these highly degenerate primers (Tables S4 and S5). *Escherichia coli* provided the negative genomic control.

Database searches

To obtain complete genomic SHC sequences and translated environmental *sqhC* fragments, parameters for TBLASTN were the reference SHC sequence for *Alicyclobacillus acidocaldarius* (GI: 2851526); standard genetic code; expectation value cut-off 10^{-1}.

Characterized species. All available protein sequences of known and putative SHCs were obtained from the Integrated Microbial Genomes database of the Joint Genomes Institute (JGI; http://img.jgi.doe.gov/pub/main.cgi) and from the National Center for Biotechnology Information as of November 2006 (NCBI; http://www.ncbi.nlm.nih.gov/). Sequences for sterol cyclases (OSCs) were as described in Pearson and colleagues (2003).

Metagenomes. Searches were performed through the NCBI BLAST portal by selecting environmental databases (env_nt) and through JGI by selecting Integrated Microbial Genomes with Microbiomes. All available databases were searched as of November 2006. The TIGR-Monterey Bay Coastal Ocean Microbial Observatory sequencing project also was searched through The Institute for Genomic Research portal (http://www.tigr.org/tdb/MBMO/MBMO.shtml). All metagenomic SHC fragments that have translated AA lengths $\geq$ 75 AA appear in Table S2.

Alignments and tree construction

Squalene-hopene cyclase sequences were aligned using CLUSTALW through the portal http://clustalw.genome.ad.jp/. Multiple alignment parameters were: gap open penalty, 13.0; gap extension penalty 0.05; BLOSUM weight matrix for proteins; indels treated as single substitutions. Two alignments were prepared for use in phylogeny calculations. The nearly full-length alignment spanned residues 1–624 of *A. acidocaldarius* SHC and was terminated at the last known catalytic AA (Y624). The alignment covers 99% (624/631) of the enzymatic sequence of *A. acidocaldarius* and similar percentages of all other known SHCs and OSCs. It was used to calculate the tree in Fig. 4A (total of 70 sequences). A partial alignment then was constructed to include environmental sequences. This short alignment spanned residues 309–535 of *A. acidocaldarius*, corresponding to the amplicon of the Td-N PCR protocol. Included in this alignment were one species per genus from the set of organisms shown in Fig. 4A; all PCR-generated sequences from LM and NCP; and only the SS, FS and IM metagenome sequence fragments that were alignable over $>\sim$50% of this region and contained at least two of the three conserved motifs SPWVDT, PDxDD or DGGWGE. This alignment was used to calculate the tree in Fig. 4B (total of 95 sequences).

Maximum likelihood trees were calculated using PHYML (Guindon and Gascuel, 2003; http://atgc.lirmm.fr/phyml/). For the full-length sequences, three trees were calculated using substitution matrices JTT (Jones *et al*., 1992), WAG (Whelan and Goldman, 2001) and Dayhoff (Dayhoff *et al*., 1978); 100 bootstrap replicates were calculated for each. The starting tree was BIONJ; optimized topologies and branch lengths were selected. For the short-length environmental sequences, only the JTT substitution matrix was calculated (100 replicates). All other parameters were the same, except that the initial tree was user-supplied with the initial topology of the genomic sequences the same as in Fig. 4A. In all calculations, the proportion of invariable sites was estimated by the program.

Rarefaction analysis

Rarefaction curves for the LM and NCP sequencing project were calculated using Distance-based OTU and Richness determination (DOTUR; Schloss and Handelsman, 2005). Distance matrices for the DNA sequences of the clones described in Table S1 were calculated using default parameters of *dnadist* in PHYLIP (http://evolution.genetics.washington.edu/phylip.html). Sequences were then assigned to OTUs based on the UPGMA (average neighbour) clustering algorithm implemented in DOTUR with default parameters for precision (0.01) and bootstrapping (1000), and rarefaction results were calculated.

Data deposition and supplemental information

The sequences reported in Table S1 have been deposited in GenBank under the Accession No. EF030657–EF030690. Details about PCR primers and thermocycler methods appear in *Supplementary material*.

Acknowledgements

We thank A.H. Knoll and R.E. Summons for helpful discussions; E.F. DeLong for discussions and editorial assistance; and K.-U. Hinrichs and an anonymous reviewer for their thorough reviews. M.C. Fisher, A.E. Ingalls, S.R. Shah, S.J. Carter, L.I. Aluwihare and R.Hansman are thanked for field assistance. This work was supported by NSF EAR-0311937 and NSF OCE-0241363 (to A.P.) and the David and Lucille Packard Foundation. Funding for genomic sequencing of selected microbial species and the availability of these data in the public domain are supported by The Institute for Genome Research (TIGR), the US Department of Energy (DOE and the J. Craig Venter Institute. We acknowledge the United Kingdom as the country of origin for the Sargasso Sea metagenome data.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Berry, A.M., Harriott, O.T., Moreau, R.A., Osman, S.F., Benson, D.R., and Jones, A.D. (1993) Hopanoid lipids compose the *Frankia* vesicle envelope, presumptive barrier of oxygen diffusion to nitrogenase. *Proc Natl Acad Sci USA* **90:** 6091–6094.

Blair, N., Leu, A., Munoz, E., Olson, J., Kwong, E., and DesMarais, D. (1985) Carbon isotopic fractionation in heterotrophic microbial metabolism. *Appl Environ Microbiol* **50:** 996–1001.

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

Brassell, S.C., Eglinton, G., Marlowe, I.T., Pflaumann, U., and Sarnthein, M. (1986) Molecular stratigraphy: a new tool for climatic assessment. *Nature* **320:** 129–133.

Brocks, J.J., Logan, G.A., Buick, R., and Summons, R.E. (1999) Archean molecular fossils and the early rise of eukaryotes. *Science* **285:** 1033–1036.

Brocks, J.J., Love, G.D., Summons, R.E., Knoll, A.H., Logan, G.A., and Bowden, S.A. (2005) Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437:** 866–870.

Conte, M.H., Weber, J.C., and Ralph, N. (1998) Episodic particle flux in the deep Sargasso Sea: an organic geochemical assessment. *Deep Sea Res I* **45:** 1819–1841.

Conte, M.H., Dickey, T.D., Weber, J.C., Johnson, R.J., and Knap, A.H. (2003) Transient physical forcing of pulsed export of bioreactive material to the deep Sargasso Sea. *Deep Sea Res I* **50:** 1157–1187.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence Structure*, Vol. 5, Suppl. 3. Dayhoff, M.O. (ed.). Washington, DC, USA: National Biomedical Research Foundation, pp. 345–352.

DeLong, E.F. (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **3:** 459–469.

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

Farrimond, P., Head, I.M., and Innes, H.E. (2000) Environmental influence on the biohopanoid composition of recent sediments. *Geochim Cosmochim Acta* **64:** 2985–2992.

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

Fischer, W.W., Summons, R.E., and Pearson, A. (2005) Targeted genomic detection of biosynthetic pathways: anaerobic production of hopanoid biomarkers by a common sedimentary microbe. *Geobiology* **3:** 33–40.

Förster, H.J., Biemann, K., Haigh, W.G., Tattrie, N.H., and Colvin, J.R. (1973) The structure of novel C35 pentacyclic terpenes from *Acetobacter xylinium*. *Biochem J* **135:** 133–143.

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W.J., Rosenthal, A., and Perret, X. (1997) Molecular basis of symbiosis between Rhizobium and legumes. *Nature* **387:** 394–401.

Gill, S.R., Pop, M., DeBoy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312:** 1355–1359.

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

Härtner, T., Straub, K.L., and Kannenberg, E. (2005) Occurrence of hopanoid lipids in anaerobic *Geobacter* species. *FEMS Microbiol Lett* **243:** 59–64.

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

Hoshino, T., Kouda, M., Abe, T., and Sato, T. (2000) Functional analysis of Phe605, a conserved aromatic amino acid in squalene-hopene cyclases. *Chem Commun* **16:** 1485–1486.

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

Ingalls, A.E., Shah, S.R., Hansman, R.L., Aluwihare, L.I., Santos, G.M., Druffel, E.R.M., and Pearson, A. (2006) Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* **103:** 6442–6447.

Innes, H.E., Bishop, A.N., Head, I.M., Farrimond, P. (1997) Preservation and diagenesis of hopanoids in recent lacustrine sediments of Priest Pot, England. *Organic Geochemistry* **26:** 565–576.

Jahnke, L.L., Summons, R.E., Hope, J.M., and Des Marais, D.J. (1999) Carbon isotopic fractionation in lipids from methanotrophic bacteria II: the effects of physiology and environmental parameters on the biosynthesis and isotopic signatures of biomarkers. *Geochim Cosmochim Acta* **63:** 79–93.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8:** 275–282.

Kleemann, G., Alskog, G., Berry, A.M., and Huss-Danell, K. (1994) Lipid composition and nitrogenase activity of the symbiotic *Frankia* (*Alnus incana*) in response to different oxygen concentrations. *Protoplasma* **183:** 107–115.

Lin, W.C., Coppi, M.V., and Lovley, D.R. (2004) *Geobacter sulfurreducens* can grow with oxygen as a terminal electron acceptor. *Appl Environ Microbiol* **70:** 2525–2528.

Mahenthiralingam, E., Baldwin, A., Drevinek, P., Vanlaere, E., Vandamme, P., Lipuma, J.J., and Dowson, C.G. (2006) multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* **1:** e17 doi:10.1371/journal.pone.0000017.

Martin, H.G., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C., *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24:** 1263–1269.

Merkofer, T., Pale-Gosdemange, C., Wendt, K.U., Rohmer, M., and Poralla, K. (1999) Altered product pattern of a squalene-hopene cyclase by mutagenesis of active site residues. *Tetrahedron Lett* **40:** 2121–2124.

Nalin, R., Putra, S.R., Domenach, A.M., Rohmer, M., Gourbiere, F., and Berry, A.M. (2000) High hopanoid/total lipids ratio in *Frankia mycelia* is not related to the nitrogen status. *Microbiology (UK)* **146:** 3013–3019.

Ochs, D., Kaletta, C., Entian, K.-D., Beck-Sickinger, A., and Poralla, K. (1992) Cloning, expression, and sequencing of a squalene-hopene cyclase, a key enzyme in triterpenoid metabolism. *J Bacteriol* **174:** 298–302.

Ourisson, G., and Albrecht, P. (1992) Hopanoids. 1. Geohopanoids: the most abundant natural products on Earth? *Acc Chem Res* **25:** 398–402.

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

Pale-Gosdemange, C., Feil, C., Rohmer, M., and Poralla, K. (1998) Occurrence of cationic intermediates and deficient control during the enzymatic cyclization of squalene to hopanoids. *Angew Chem Int Ed Engl* **37:** 2237–2240.

Pancost, R.D., Sinninghe Damsté, J.S., de Lint, S., van der Maarel, M., and Gottschal, J.C. (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., Brocks, J.J., and Budin, M. (2003) Phylogenetic and biochemical evidence for sterol synthesis in the bacterium, *Gemmata obscuriglobus*. *Proc Natl Acad Sci USA* **100:** 15352–15357.

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

Peters, K.E., Walters, C.C., and Moldowan, J.M. (2005) *The Biomarker Guide*, edn 2. Cambridge, UK: Cambridge University Press.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64:** 3724–3730.

Poralla, K., Muth, G., and Hartner, T. (2000) Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). *FEMS Microbiol Lett* **189:** 93–95.

Prahl, F.G., and Wakeham, S.G. (1987) Calibration of unsaturation patterns in long-chain ketone compositions for paleotemperature assessment. *Nature* **330:** 367–369.

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

Rohmer, M., Bouvier, P., and Ourisson, G. (1979) Molecular evolution of biomembranes: structural equivalents and phylogenetic precursors of sterols. *Proc Natl Acad Sci USA* **76:** 847–851.

Rohmer, M., Bouvier-Nave, P., and Ourisson, G. (1984) Distribution of hopanoid triterpenes in Prokaryotes. *J Gen Microbiol* **130:** 1137–1150.

Sawai, S., Akashi, T., Sakurai, N., Suzuki, H., Shibata, D., Ayabe, S.I., and Aoki, T. (2006) Plant lanosterol synthase: divergence of the sterol and triterpene biosynthetic pathways in eukaryotes. *Plant Cell Physiol* **47:** 673–677.

Schloss, P.D., and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71:** 1501–1506.

Schouten, S., Hopmans, E.C., Schefuß, E., and Sinninghe Damsté, J.S. (2002) Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth Planet Sci Lett* **204:** 265–274.

Sinninghé Damste, J.S., Muyzer, G., Abbas, B., Rampen, S.W., Masse, G., Allard, W.G., *et al.* (2004a) The rise of the rhizosolenoid diatoms. *Science* **304:** 584–587.

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

Summons, R.E., and Jahnke, L.L. (1992) Hopenes and hopanes methylated in ring-A: correlation of the hopanoids from extant methylotrophic bacteria with their fossil analogues. In *Biological Markers in Sediments and Petroleum.* Moldowan, J.M., Albrecht, P., and Philp, R.P. (eds). Englewood Cliffs, NJ, USA: Prentice Hall, pp. 182–200.

Summons, R.E., Jahnke, L.L., Logan, G.A., and Hope, J.M. (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400:** 554–557.

Summons, R.E., Bradley, A.S., Jahnke, L.L., and Waldbauer, J.R. (2006) Steroids, triterpenoids and molecular oxygen. *Phil Trans Royal Soc B* **361:** 951–968.

Talbot, H.M., Watson, D.F., Murrell, J.C., Carter, J.F., and Farrimond, P. (2001) Analysis of intact bacteriohopanepolyols from methanotrophic bacteria by reversed-phase high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometery. *J Chromatogr A* **921:** 175–185.

Thiel, V., Peckmann, J., Richnow, H.H., Luth, U., Reitner, J., and 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.

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

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., and Chang, H.W., *et al.* (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43.

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

Wendt, K.U., Poralla, K., and Schultz, G.E. (1997) Structure and function of a squalene cyclase. *Science* **277:** 1811–1815.

Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18:** 691–699.

Wick, L.Y., McNeill, K., Rojo, M., Medilanski, E., and Gschwend, P.M. (2000) Fate of benzene in a Stratified Lake receiving contaminated groundwater discharges from a Superfund site. *Environ Sci Technol* **34:** 4354–4362.

Xie, S.C., Pancost, R.D., Yin, H.F., Wang, H.M., and Evershed, R.P. (2005) Two episodes of microbial change coupled with Permo/Triassic faunal mass extinction. *Nature* **434:** 494–497.

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

Supplementary material

The following supplementary material is available for this article online:

Additional Methods. Description of PCR, cloning and sequencing protocols.

Fig. S1. Pictorial representation of sample–primer–program combinations. Samples TJ1 and TJ2 are from LM; sample HIS is from the NCP.

Fig. S2. Sample TJ1. Amplification of a 900 bp region of *sqhC* using fully degenerate primers designed to the cyanobacterial/actinomycetes group. Primers SHCF, SHCcyR. M, DNA marker; (–), template containing no SHC; (+), sample TJ1.

Fig. S3. Sample TJ1. Amplification of *sqhC* using touchdown-nested programs. Primers SHCF, SHCcyR; amplicon 900 bp. Second amplification of this sample using nested primers SHCF and SHCNR; intended amplicon 700 bp and truncated amplicon 350 bp. M, DNA marker; Td, touchdown amplicon; N, nested amplicon.

Table S1. Data for LM and NCP clones. Columns: (1) clone; (2) letter codes corresponding to Fig. 2; (3) closest described genus by TBLASTN; (4–8) expectation values and AA statistics; (9) other clones obtained by PCR with ≥ 99% translated AA identity to the named clone; (10) total frequency data used to create Fig. 2A and C; (11,12) confirmation of conserved motifs expected for SHCs (p., primer region).

Table S2. Data for environmental metagenomes SS, FS, IM. Columns: (1) name; (2) GenBank GI number; (3) sequence coverage type; (4) closest described genus by TBLASTN; (5–8) expectation values and AA statistics.

Table S3. Summary of the eight combinations of primer sets and PCR programs attempted for samples TJ1, TJ2 and HISurf. Symbols indicate (+) sequences obtained; (–) attempted, but no *sqhC* sequences obtained; *n.a.*, no amplicon; shaded, not attempted.

Table S4. Genera matched by forward and reverse primers, based on bacterial genomes available through JGI.

Table S5. Translation of each degenerate primer, with melting temperatures.

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

