The ISME Journal (2014) 8, 675–684

www.nature.com/ismej

ORIGINAL ARTICLE

Diverse capacity for 2-methylhopanoid production correlates with a specific ecological niche

Jessica N Ricci[1], Maureen L Coleman[1,6], Paula V Welander[2,7], Alex L Sessions[3], Roger E Summons[2], John R Spear[4] and Dianne K Newman[1,3,5]

[1]*Division of Biology, California Institute of Technology, MC156-29, 1200 E. California Boulevard, Pasadena, CA 91125, USA;* [2]*Department of Earth, Atmospheric and Planetary Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E25-633, Cambridge, MA 02139, USA;* [3]*Division of Geological and Planetary Sciences, California Institute of Technology, MC100-23, 1200 E. California Boulevard, Pasadena, CA 91125, USA;* [4]*Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO 80401, USA and* [5]*Howard Hughes Medical Institute, MC156-29, 1200 E. California Boulevard, Pasadena, CA 91125, USA*

Molecular fossils of 2-methylhopanoids are prominent biomarkers in modern and ancient sediments that have been used as proxies for cyanobacteria and their main metabolism, oxygenic photosynthesis. However, substantial culture and genomic-based evidence now indicates that organisms other than cyanobacteria can make 2-methylhopanoids. Because few data directly address which organisms produce 2-methylhopanoids in the environment, we used metagenomic and clone library methods to determine the environmental diversity of *hpnP*, the gene encoding the C-2 hopanoid methylase. Here we show that *hpnP* copies from alphaproteobacteria and as yet uncultured organisms are found in diverse modern environments, including some modern habitats representative of those preserved in the rock record. In contrast, cyanobacterial *hpnP* genes are rarer and tend to be localized to specific habitats. To move beyond understanding the taxonomic distribution of environmental 2-methylhopanoid producers, we asked whether *hpnP* presence might track with particular variables. We found *hpnP* to be significantly correlated with organisms, metabolisms and environments known to support plant–microbe interactions (*P*-value $< 10^{-6}$); in addition, we observed diverse *hpnP* types in closely packed microbial communities from other environments, including stromatolites, hot springs and hypersaline microbial mats. The common features of these niches indicate that 2-methylhopanoids are enriched in sessile microbial communities inhabiting environments low in oxygen and fixed nitrogen with high osmolarity. Our results support the earlier conclusion that 2-methylhopanoids are not reliable biomarkers for cyanobacteria or any other taxonomic group, and raise the new hypothesis that, instead, they are indicators of a specific environmental niche.

The ISME Journal (2014) **8,** 675–684; doi:10.1038/ismej.2013.191; published online 24 October 2013
Subject Category: Geomicrobiology and microbial contributions to geochemical cycles
Keywords: biomarker; 2-methylhopanoid; plant–microbe interaction

Introduction

Morphological and molecular fossils left by microorganisms in ancient sedimentary rocks can provide a valuable window into the early history of life on Earth. Yet due to challenges inherent in working with billion-year-old samples, the interpretation of these fossils has often been contentious (Schopf and Packer 1987; Walter *et al.,* 1992; Brasier *et al.,* 2002, 2004). In this context, organic biomarkers have received attention due to their potential to provide more specific information about the composition of ancient microbial communities (Brocks and Pearson, 2005). Hopanes and steranes are among the more prominent classes of these biomarkers (Rohmer, 2010). These molecules can unambiguously be interpreted as the diagenetic remains of hopanoids and steroids, polycyclic triterpenoids found in the membranes of numerous organisms today (Rohmer *et al.,* 1984; Ourisson *et al.,* 1987). However, ambiguity regarding their

Correspondence: Professor DK Newman, Division of Biology, California Institute of Technology, MC156-29, 1200 E. California Boulevard, Pasadena, CA 91125, USA.
E-mail: dkn@caltech.edu
[6]Present address: Department of the Geophysical Sciences, University of Chicago, 5734 S. Ellis Avenue, Chicago, IL 60637
[7]Present address: Department of Environmental Earth System Science, Stanford University, 473 Via Ortega Road, Rm 140, Stanford, CA 94305
Received 14 July 2013; revised 4 September 2013; accepted 22 September 2013; published online 24 October 2013

distribution and function in modern bacteria clouds our ability to interpret their fossils in the rock record.

Hopanoids structurally resemble steroids, but unlike steroids, they are primarily made by bacteria and do not require oxygen for their biosynthesis (Ourisson *et al.,* 1987; Fischer *et al.,* 2005; Rashby *et al.,* 2007). Although hopanoids exhibit structural diversity of their side chains, much of this diversity is lost through diagenesis, resulting mainly in the preservation of their hydrocarbon skeletons, hopanes. Important exceptions to this are methyl groups at C-2 or C-3, which can be preserved. In fact, 2-methylhopanes have a rich history in the fossil record, found at discrete times and locations as far back as 2.7 billion years (Brocks *et al.,* 1999), although this latter finding is under scrutiny (Rasmussen *et al.,* 2008). The varied distribution of 2-methylhopanes in the more 'recent' rock record (that is, million-year timescales), showing peaks in abundance correlated with ocean anoxic events, suggests their production may be linked to particular environmental triggers (Knoll *et al.,* 2007).

Until recently, 2-methylhopanes were viewed as biomarkers for cyanobacteria and their main energy-generating metabolism, oxygenic photosynthesis (Summons *et al.,* 1999). However, the finding of conditional 2-methylhopanoid (2-MeBHP) production by the anoxygenic phototroph *Rhodopseudomonas palustris* called this interpretation into question (Rashby *et al.,* 2007). From genomic data and culture-based work, it is now clear that only a minority of cyanobacteria (that is, 13% of all sequenced cyanobacterial species and 19% of all sequenced cyanobacterial genera) have the gene, *hpnP*, that encodes the enzyme responsible for methylating hopanoids at C-2 (Supplementary Figure S1); (Talbot *et al.,* 2008; Welander *et al.,* 2010). Moreover, other bacterial species possess *hpnP* including many from a subclade of alphaproteobacteria and an acidobacterium (Talbot *et al.,* 2007a; Welander *et al.,* 2010).

Despite the known distribution of *hpnP* among cultured organisms, uncertainty remains about which producers of 2-MeBHPs are environmentally relevant and whether there is a common ecology that correlates with the production of these molecules. Given that 2-MeBHPs are produced by diverse bacteria, we asked two questions: (1) what is the potential for 2-MeBHP production by different organisms in modern environments, and (2) might 2-MeBHP producers inhabit common ecological niche(s)? Here we assess the distribution of *hpnP* in various environments, and find a statistically significant correlation with modern habitats that support plant–microbe interactions. We discuss the implications of these results for interpreting the ancient 2-methylhopane record.

Materials and methods

Distribution of hpnP in cyanobacterial genomes
The abundance of *hpnP* genes among cyanobacteria was calculated using finished cyanobacterial genomes on the Joint Genomes Institute's Integrated Microbial Genomes (IMG) database (Markowitz *et al.,* 2011). Additionally, we analyzed this data set condensed to the genus level to reduce bias (Supplementary Figure S1).

Distribution of hpnP in metagenomes
To retrieve HpnP sequences from public metagenomes, HpnP from *R. palustris* TIE-1 (NC_011004.1) was used as a query against the NCBI metagenomic proteins database, IMG/M, CAMERA and myMGDB. All hits with an *e*-value equal to or less than 1×10^{-50} were subjected to phylogenetic analysis (Markowitz *et al.,* 2011; Sun *et al.,* 2011). Approximately 20% of hits retrieved clustered phylogenetically with known HpnP sequences, whereas all others clustered with sequences known not to be *hpnP*. Only sequences that cluster with known *hpnP* genes were identified as *hpnP* (Supplementary Data Set S1). All searches were completed in December 2012. Descriptions of surveyed metagenomes appear in Supplementary Data Set S2.

Sequences were identified as *gyrB*, *psbC* and *shc* if they had an *e*-value equal to or less than 1×10^{-20}. *Escherichia coli gyrB* (NC_000913.2), *Nostoc punctiforme psbC* (YP_001866969.1) and *R. palustris* TIE-1 *shc* (NC_011004.1) were used as query sequences. This *e*-value was used because it captured known diversity of the genes without retrieving related sequences with other functions. Sequences of *shc* were determined to be cyanobacterial or alphaproteobacterial by top BLASTP hit (Table 1).

Clone library and sample preparation
DNA samples from Yellowstone National Park hot springs were collected and prepared as previously reported (Osburn *et al.*, 2011). All other samples were extracted with the UltraClean Soil DNA Isolation Kit (MoBio, Carlsbad, CA, USA). DNA samples were stored at $-20\,^{\circ}C$. Information on samples can be found in Supplementary Table 1.

Nested PCR primers for *hpnP* were designed based on the conserved amino-acid motifs (A/G)FMPPQ and (S/T)GII(L/M)G, for the first pair, and (A/V)(L/I)GGPS and GIETP(E/D), for the second. The resulting primers were hpnP_1 F 5′-GSB TTY ATG CCD CCB CAR GG-3′, hpnP_1 R 5′-TCN ARK CCV AKR ATR ATN CC-3′, hpnP_2 F 5′-GYB VTB GGH GGN CCN TCN GT-3′ and hpnP_2 R 5′-TCN GGN GTY TCD ATN CC-3′, respectively (Supplementary Figure S3A). To amplify *hpnP*, Promega PCR master mix (Promega, Madison, WI, USA) and cycles of 95 °C for 2 min, 95 °C denaturation for 30 s, 50 °C for 30 s and 72 °C for 1 min for 35

Table 1 Abundances of *hpnP* and *shc* in metagenomes

Environments[a]	*Mbp*	*Reads*	*gyrB*[b]	*psbC*[c]	*hpnP*					*shc*		
					Total	*Cn*	*Al*	*Ac*	*Un*	*Total*	*Cn*	*Al*
Hot springs	2543	6852165	1015	73	3	3	—	—	—	48	2	3
Terrestrial	44 645	272 985 945	8562	126	87	3	74	—	10	1304	27	610
Soil and rhizosphere	41 856	266 263 412	6037	125	80	3	68	—	9	1119	26	514
Insect fungal gardens	2376	5 735 261	2197	1	6	—	5	—	1	146	—	74
Wood compost	413	987 272	328	—	1	—	1	—	—	39	1	22
Freshwater	58 136	305 595 686	8810	200	21	—	10	—	11	554	16	49
Lentic and lotic	6852	17 241 435	5674	190	17	—	8	—	9	423	14	32
Groundwater	51 172	288 125 962	2544	8	2	—	—	—	2	115	2	13
Wastewater	111	228 289	592	2	2	—	2	—	—	16	—	4
Marine	203 016	4 793 195 841	13 388	1759	28	—	26	—	2	488	4	192
Open ocean	4509	4 359 663	1142	534	—	—	—	—	—	48[d]	—[d]	24[d]
Coastal, upwelling, harbor	188 571	4 767 074 572	7805	491	2	—	—	—	2	137[d]	1[d]	37[d]
Estuary	2971	6 629 924	2405	10	9	—	9	—	—	183[d]	—[d]	87[d]
High latitude	1888	4 405 566	526	7	—	—	—	—	—	17	—	4
Ace Lake	1119	2 465 421	20	506	2	—	2	—	—	9	—	8
Deep sea hydrothermal vent	1221	2 197 778	424	3	15	—	15	—	—	14	—	4
Reef	627	597 699	258	92	—	—	—	—	—	1[d]	—	—[d]
Mangrove	153	148 018	44	1	—	—	—	—	—	5[d]	—[d]	—[d]
Hypersaline	1411	3 260 701	490	50	—	—	—	—	—	71[d]	1[d]	27[d]
Equatorial upwelling	163	712 274	102	30	—	—	—	—	—	—	—	—
Spring bloom	223	1 064 091	102	13	—	—	—	—	—	1	—	1
Trichodesmium bloom	159	280 134	70	22	—	—	—	—	—	2	2	—

Abbreviations: Cn, Cyanobacterial; Al, Alphaproteobacterial; Ac, Acidobacterial; Un Unknown.
[a]Identifying information and descriptions of metagenomes included here appear in Supplementary Data Set S2.
[b]*gyrB*, estimates the number of bacterial genomes.
[c]*psbC*, estimates the number of cyanobacterial genomes.
[d]Some data from Pearson and Rusch (2009).

cycles, followed by 72 °C for 3 min were used. An aliquot of the first PCR (1 μl) was used as a template for the second. We validated this method by testing the primer sets on diverse 2-MeBHP-producing organisms as well as some that do not make 2-MeBHPs (Supplementary Figure S3C).

PCR products from the second reaction were extracted with the Montage gel extraction kit (Millipore, Billerica, MA, USA) and cloned with the TOPO TA Cloning Kit for sequencing using One Shot Top10 electrocompetent cells (Invitrogen, Carlsbad, CA, USA). Twenty-four to 100 clones per sample were amplified by PCR and restriction digested with AluI (New England Biolabs, Ipswich, MA, USA). Amplicons with unique digestion patterns were sequenced (Retrogen, San Diego, CA, USA). Sequences were trimmed to remove contaminating vector and poor quality regions, and then translated in Geneious 5.6.5. Representative sequences of 95% identity clusters were picked using CD-HIT and used to make phylogenetic trees (Huang *et al.,* 2010). Sequences have been deposited in GenBank under the accession numbers KC603770 thru KC603846.

Rarefaction curves

The number of unique restriction digestion patterns was used as a proxy for *hpnP* diversity. To generate rarefaction curves, the species observed metric was used in the alpha diversity package of QIIME 1.5.0 (Caporaso *et al.,* 2010). Data sets were rarefied 100 times in 2 step increments (Supplementary Figure S5). To compare richness of *hpnP* between clone libraries, data sets were rarefied to 25 clones to avoid depth of sampling bias and averaged (Supplementary Figure S5). Guerrero Negro sample 4 was not included in this analysis because of low sampling coverage (Supplementary Table 1).

Alignment and phylogeny construction

All alignments were made using the MAFFT v6.859b l-ins-i algorithm (Katoh and Toh, 2008). Reference HpnP alignments comprise all HpnP sequences retrieved from NCBI as of December 2012 and *hpnP* from *Phormidium luridum* UTEX 426. Full-length *P. luridum* UTEX 426 *hpnP* was obtained by inverse PCR using the degenerate *hpnP* PCR primers as a probe. Outgroup sequences were picked from the sister clade of HpnP (Welander *et al.,* 2010). The reference alignment was trimmed in Gblocks 0.91b with relaxed parameters (Talavera and Castresana, 2007). Environmental HpnP sequences were then added to the reference HpnP alignments using the seed option in MAFFT. Phylogenetic trees were made by PhyML v3.0 using the LG model with aLRT supports and modified in iTOL (Guindon *et al.,* 2010; Letunic and Bork, 2007).

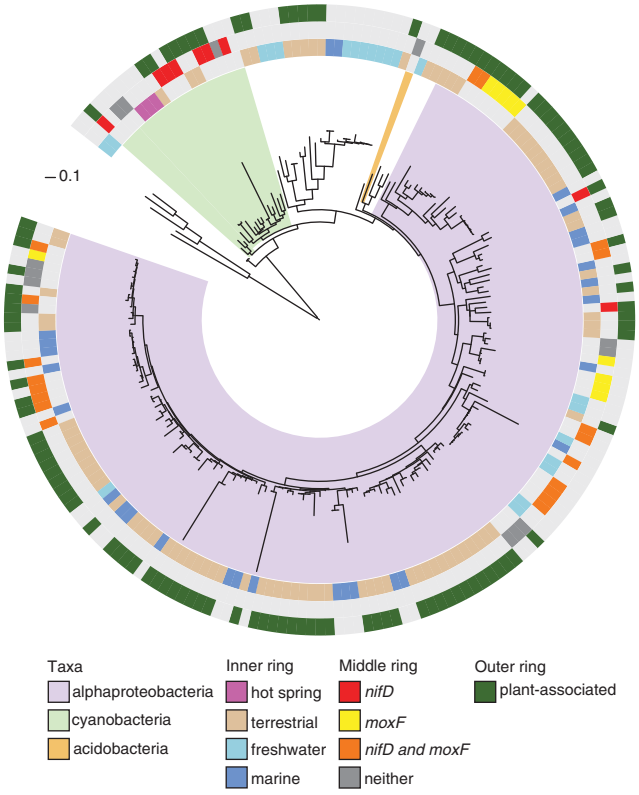

Figure 1 HpnP diversity from metagenomes and its correlation with plant–microbe interactions. Metagenomic databases were searched for *hpnP*-like sequences. Sequences that could be phylogenetically classified as members of the HpnP family appear in this maximum likelihood phylogeny with HpnP sequences from genomes. The colored ranges on the tree's branches indicate alphaproteobacterial, cyanobacterial or acidobacterial clades of HpnP determined by HpnP sequences from reference genomes (Supplementary Figure S2). Metagenomic sequences that fall within one of these clades are classified as belonging to that taxon. In the inner ring surrounding the tree, HpnP sequences from metagenomes are colored by environment of origin, either hot spring, terrestrial, freshwater or marine. In the middle ring, HpnP sequences from genomes are colored to indicate the presence of *nifD*, *moxF* or both genes in the same genome. The outer ring indicates that the organism the sequence derives from was isolated from a plant-associated environment, has an established plant interaction or that a metagenomic sequence is from a plant-associated environment (soil or rhizosphere, insect waste dumps, wood compost). The light gray background corresponds to no data in the inner two rings (for example, genomes do not have an environment of origin color) or no plant association found in the outer ring; the outgroup was not included in the analysis. aLRT support values and leaf names are shown in Supplementary Figure S2. The scale bar is a measure of evolutionary distance equaling 0.1 substitutions per site.

HpnP types (cyanobacterial, alphaproteobacterial or unknown) from metagenomic and clone library sequences were classified as such when they grouped phylogenetically with reference HpnP sequences of the same type (Figure 1; Supplementary Figures S2 and S4). Unknown *hpnP* groups were defined by a lack of reference sequences. Metagenomic *hpnP* hits with long branches were examined for recombination events using Recombination Analysis Tool, but none were found (Etherington *et al.,* 2005).

hpnP correlation with plants

The initial observation that many *hpnP*-containing organisms and metagenomes were plant associated used the following criteria for a positive plant association: the organism was isolated from a plant-associated environment, had a known plant interaction established in the literature or the metagenome was from a plant-derived environment (that is, soil and rhizosphere, wood compost and insect fungal gardens, as these are maintained by leaf-cutting ants) based on available metadata (Figure 1 outer ring).

To assess the significance of the relationship between *hpnP* or *shc* and plant-associated organisms or environments, we used the hypergeometric test to evaluate if there was non-random overlap between organisms or environments that have *hpnP* or *shc* and those that are plant associated (Supplementary Table 2). Plant association for metagenomes was determined as described above. When addressing this analysis among organisms, we included in our analysis all finished bacterial genomes condensed to the species level on IMG. Organisms were counted as plant associated if a plant species was listed under 'host name' in the description of the genome. As filling out this data field is voluntary, we would expect more false negatives than false positives. In an attempt to reduce the number of false negatives, we mined Pubmed for abstracts describing plant associations among our list of organisms. To conduct an unbiased search, we used the following Boolean expression: two-word name of the species AND host plant OR plant host OR plant-microbe OR root-coloniz* OR plant-associat* NOT pathogen, where * allows for multiple endings. Abstract hits were manually annotated and positive hits were combined with the plant-associated list from IMG.

The hypergeometric test was also used to assess an *hpnP* correlation with *nifD* and *moxF* in finished bacterial genomes from IMG (Supplementary Table 3). Genomes were found to have *nifD or moxF* if they returned an *e*-value less than or equal to 1×10^{-50} when using *Nostoc* sp. PCC 7120 *nifD* (gi 4376092) and *Methylobacterium extorquens* AM1 *moxF* (YP_002965446) as queries. Similar results for *nifD* were also obtained for *nifH*.

Hopanoid analysis

Select samples were targeted for analysis by liquid chromatography-mass spectrometry for a limited number of hopanoids. These samples appear in Figure 2 with either an identifier, if hopanoids were analyzed, or blank, if hopanoids were not analyzed. We also attempted to quantify hopanoids, specifically unextended hopanoids that cannot be identified by liquid chromatography-mass spectrometry, using standard GC-MS techniques (Sessions *et al.,* 2013), but we were unable to unambiguously detect hopanoids due to a high background. Samples

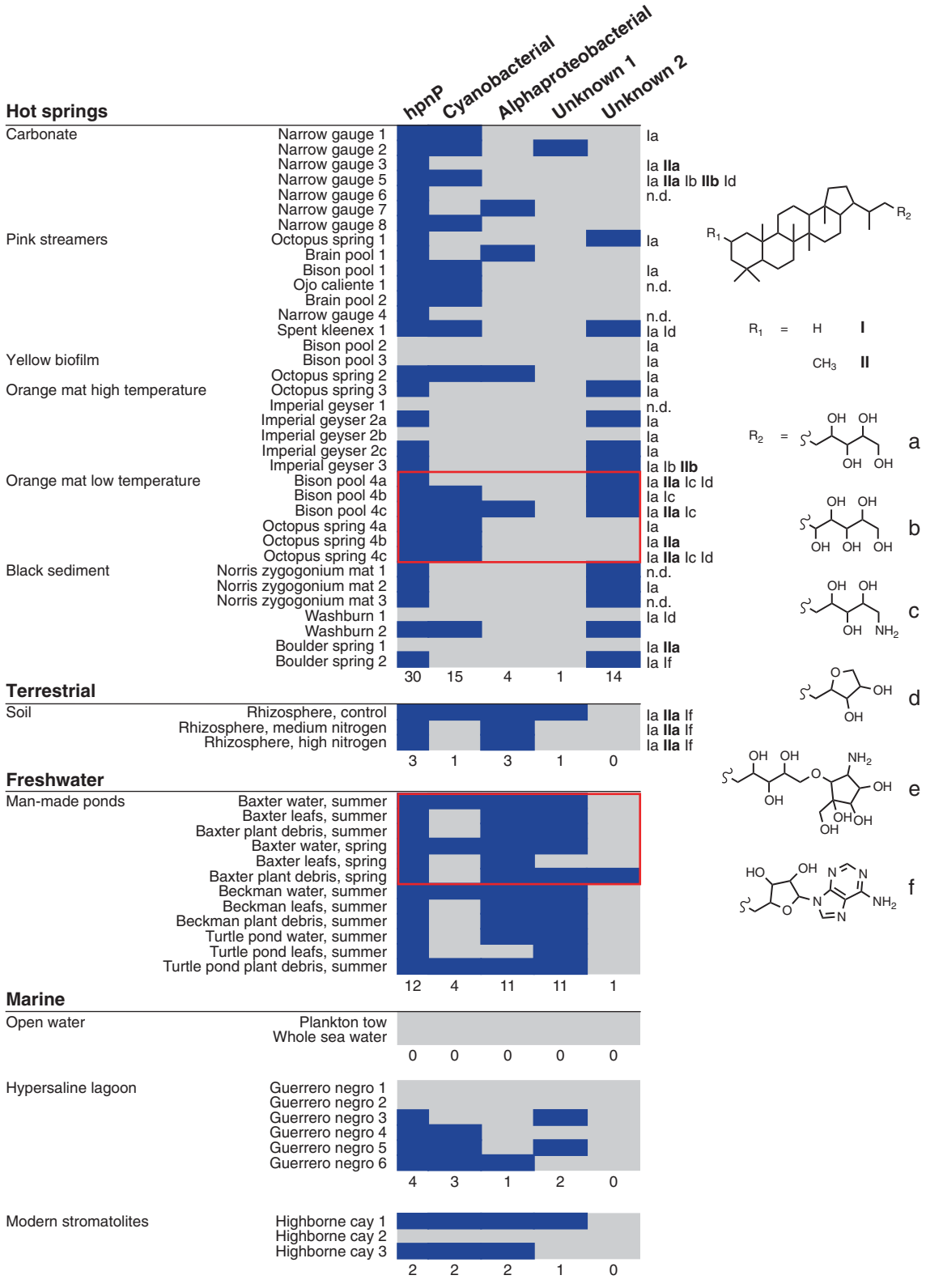

Figure 2 Distribution of HpnP types in clone libraries and hopanoids from various environments. HpnP sequences generated from each sample were classified by HpnP type in a phylogenetic tree (Supplementary Figure S4). The presence of *hpnP* or an *hpnP* type is indicated in blue. Total abundances of *hpnP* types are reported below each environment. A red outline highlights spatial and temporal heterogeneity. Sample descriptions appear in Supplementary Table 1. Presence of hopanoids was determined for select samples by LC-MS. Abbreviations: Ia, bacteriohopanetetrol; IIa, 2-methylbacteriohopanetetrol; Ib, bacteriohopanepentol; IIb, 2-methylbacteriohopanepentol; Ic, aminobacteriohopnetriol; Id, anhydrobacteriohopanetetrol; Ie, adenosylhopane; and n.d., not detected. Other functionalized hopanoids, including aminohopanoids, were not detected. Methylated hopanoids (IIa and IIb) are highlighted in bold type.

were extracted as previously reported (Sessions *et al.,* 2013).

Methylated and non-methylated bacteriohopanepolyols were identified by liquid chromatography-mass spectrometry as previously described (Welander *et al.,* 2012). Lipids were acetylated by incubating total lipid extract (1 mg) from each sample in a 1:1 (v:v, 250 μl) mixture of acetic anhydride (Sigma-Aldrich, St. Louis, MO, USA) and pyridine (Sigma-Aldrich) for 1 h at 70 °C. Acetylated total lipid extracts were dried down under a stream of N_2 and resuspended in methanol (1 ml) for a final total lipid extract concentration of 1 μg μl^{-1}. Subsequently, each sample (5 μl) were loaded onto the liquid chromatography-mass spectrometry for analysis, a 1200 Series HPLC (Agilent Technologies, Santa Clara, CA, USA) equipped with an autosampler and a binary pump linked to a Q-TOF 6520 mass spectrometer (Agilent Technologies) via an atmospheric pressure chemical ionization interface (Agilent Technologies) operated in positive ion mode. The analytical procedure was adapted from (Talbot *et al.,* 2001). A Poroshell 120 EC-C18 column (2.1 × 150 mm, 2.7 μm; Agilent Technologies), set at 30 °C, was eluted isocratically first with MeOH/water (95:5, v-v) for 2 min at a flow rate of 0.15 ml min^{-1}, then using a linear gradient up to 20% (v) of isopropyl alcohol over 18 min at a flow rate of 0.19 ml min^{-1}, and isocratic for 10 min. The linear gradient was then set to 30% (v) of isopropyl alcohol at 0.19 ml min^{-1} over 10 min and maintained for 5 min. The column was subsequently eluted using a linear gradient up to 80% isopropyl alcohol (v) over 1 min at a flow rate of 0.15 ml min^{-1} and isocratic for 14 min. Finally, the column was eluted with MeOH/water (95:5, v-v) at 0.15 ml min^{-1} for 5 min. The atmospheric pressure chemical ionization parameters were as follows: gas temperature 325 °C, vaporizer temperature 350 °C, drying gas (N_2) flow 6 l min^{-1}, nebulizer (N2) flow 30 l min^{-1}, capillary voltage 1200 V, corona needle 4 μA, fragmentor 150 V. Data were recorded by scanning from m/z 100–1600. Bacteriohopanepolyols were identified on the basis of accurate mass measurements of their protonated molecular ions, fragmentation patterns in MS-MS mode and by comparison of relative retention time and the mass spectra with published data (Talbot, Squier, *et al.,* 2003; Talbot, Summons, *et al.,* 2003; Talbot *et al.,* 2007b).

Results and discussion

Environmental distribution of hpnP

To identify potential biological sources of environmental 2-MeBHPs, we followed a two-pronged approach: (1) we searched all available metagenomes for the presence of *hpnP* (Figure 1; Table 1; Supplementary Figure S2) and (2) we generated clone libraries of *hpnP* sequences from diverse environments (Figure 2; Supplementary Table 1). In the first survey, 59 metagenomes were found to have *hpnP*, which resulted in the identification of 139 partial *hpnP* sequences (Supplementary Figure S2 and Supplementary Data Set S1). We also searched the same metagenomes for *shc*, which encodes squalene hopene cyclase, the enzyme that catalyzes the first step in hopanoid biosynthesis. For the second approach, we designed degenerated PCR primers that amplify all known diversity of *hpnP* (Supplementary Figure S3) and were used to retrieve 76 unique *hpnP* sequences from 62 samples (Supplementary Figure S4). Due to the potential for PCR bias, we did not infer the abundance of any particular *hpnP* type within a given library. However, we did estimate the abundance of a particular *hpnP* type in an environment by counting the number of samples from that environment that contained that specific *hpnP* type (Figure 2). Although we cannot exclude the possibility that horizontal gene transfer confounds our taxonomic assignment of *hpnP* sequences, there is no evidence of recent transfer events in the phylogeny of *hpnP* (Welander *et al.*, 2010).

Based on both metagenomic and clone library data, we found that cyanobacterial *hpnP* copies are not ubiquitous in most modern habitats, constituting only 4% of metagenomic *hpnP* sequences (Figure 2; Table 1). Consistent with this finding, we found low abundances of metagenomic cyanobacterial *shc* sequences in all environments as seen previously (Table 1); (Pearson *et al.,* 2007, 2009; Pearson and Rusch 2009). These data suggest that not only are cyanobacteria minor producers of 2-MeBHPs, but that they do not contribute substantially to hopanoid production in general. However, we cannot rule out the possibility that rare members of a community may be disproportionately active hopanoid producers.

Furthermore, some environmental *hpnP* sequences could not be identified as being from one of the known clades of 2-MeBHP produces. These sequences, which can be confidently identified as *hpnP* by homology, may represent new taxa of *hpnP*-containing organisms; alternatively, as the database of *hpnP* sequences from cultured organisms grows, we may be able to assign these sequences to previously identified clades (Figure 1, Supplementary Figures S2 and S4). Among clone library samples, the two unknown groups of *hpnP* segregate by environment of origin, with unknown group 1 found in most environments and unknown group 2 primarily in hot springs. Fourteen out of 36 hot spring clone libraries contained unknown *hpnP* sequences from group 2, nearly identical to the abundance of cyanobacterial *hpnP* sequences, 15 out of 36. Additionally, *hpnP* sequences belonging to unknown group 1 were found in 11 out of 12 pond samples (Figure 2). These observations indicate that potential novel groups of 2-MeBHP-producing bacteria may have a major role in 2-MeBHP production in hot springs and freshwater environments.

Acknowledging that comparative metagenomic analysis can be biased by the samples that have been sequenced and the methods used to sequence them, we nevertheless found a robust pattern in the available data: the majority (63%) of metagenomic *hpnP* sequences belong to terrestrial habitats, such as soil, the rhizosphere, insect fungal gardens and wood compost (Figure 1). This finding is not due to deeper sequencing of terrestrial environments. Using the abundance of *gyrB* to approximate the number of bacteria, because it is found in single copy in all bacterial genomes, we estimate that ~1% of terrestrial bacteria have *hpnP* whereas less than 0.4% of bacteria in other environments surveyed have *hpnP* (Table 1); (Biers *et al.,* 2009). In comparing *hpnP* richness between clone library data sets, rhizosphere and pond samples were found to contain more unique *hpnP* genes than the other environments tested (Supplementary Figure S5). Interestingly, 15% of terrestrial bacteria contain *shc*, whereas other habitats contain less than 7% of bacteria with *shc* (Table 1), implying that the same argument may be true for hopanoids in general. Taken together, these data suggest that terrestrial and freshwater environments harbor the majority of *hpnP* and *shc* diversity in modern ecosystems and should be considered likely sources of modern and possibly ancient (2-methyl)hopanoids. This finding is congruent with the presence of (2-methyl)hopanoids in terrestrial and freshwater settings (Talbot and Farrimond, 2007; Cooke *et al.,* 2008; Pearson *et al.,* 2009; Xu *et al.,* 2009) and observations across a land–sea transect where (2-methyl)hopanoids appeared to be partially terrestrial in origin (Sáenz *et al.,* 2011).

Paradoxically, the sedimentary context of the 2-methylhopane fossil record suggests ancient 2-MeBHP producers inhabited shallow, tropical marine environments (Knoll *et al.,* 2007). It is noteworthy that our culture-independent search for *hpnP* sequences in all marine metagenomes, highly biased towards open water, coastal and estuarine samples, did not identify any cyanobacterial *hpnP* sequences. This was not due to a lack of cyanobacteria, as 13% of bacteria in marine metagenomes were estimated to be cyanobacteria based on the abundance of *psbC,* which encodes a component of the photosynthetic machinery present at one copy per cyanobacterial genome (Table 1); (Mulkidjanian *et al.,* 2006). Of those *hpnP* sequences identified, 93% were alphaproteobacterial and 7% were of unknown origin. These sequences derived from coastal waters, estuary sediments, Ace Lake (Antarctica) and deep-sea hydrothermal vents (Table 1). These data suggest that 2-MeBHPs in some depositionally relevant modern marine environments (that is coastal water and estuaries) most likely derive from alphaproteobacteria.

To assess the capacity for 2-MeBHP production in other marine habitats that are preserved in the rock record, we analyzed *hpnP* clone libraries from Guerrero Negro hypersaline mats and Highborne Cay stromatolites (Hoehler *et al.,* 2001; Dupraz and Visscher, 2005). Among Guerrero Negro samples that contained *hpnP*, three out of four had cyanobacterial *hpnP* genes whereas an equal number also contained alphaproteobacterial or unknown *hpnP* copies. Similarly, in Highborne Cay, within the two samples found to have *hpnP*, both cyanobacterial and alphaproteobacterial copies of the gene were present (Figure 2). Although it is interesting that these habitats are the only marine environments where we find cyanobacterial *hpnP* genes, 2-MeBHP production cannot definitely be attributed to cyanobacteria in these environments. This result contrasts with a study of *hpnP* diversity in Hamelin Pool Shark Bay, Australia, another locality for stromatolite deposition, where mainly cyanobacterial *hpnP* was recovered (Garby *et al.,* 2012). It remains to be determined if this inconsistency is due to an inherent difference between these locales or due to differences in sampling and methodology; for example, the *hpnP* PCR primers used in this study were on average more degenerate than those used in Garby *et al.* (2012).

In microbial ecology, it is well known that microheterogeneity can exist over small spatial scales (Hunt *et al.,* 2008). In assessing our data, we observed spatial and temporal differences in *hpnP* diversity (Figure 2). These differences were evident even among samples with similar geochemistry, but the causes of variation are unknown. Related to this but at the level of lipid production, the presence of 2-MeBHPs did not always correlate with the presence of *hpnP* (Figure 2). Although the lack of 2-MeBHPs may have been due to the detection limit of our analysis, two additional factors may explain this observation: first, 2-MeBHP production can depend on the growth condition (Rashby *et al.,* 2007); second, the presence of 2-MeBHPs, but not *hpnP*, may have resulted from bacteria no longer present assuming the degradation rate of hopanoids is considerably slower than the disappearance of DNA. Therefore, the presence of *hpnP* does not require that a community is making 2-MeBHPs but indicates that it has the capacity to do so when the environment calls for it.

hpnP is correlated with plants

Using two independent methods, we have shown that diverse microbes have the genetic potential to produce 2-MeBHPs in a multitude of modern environments tested, making it difficult to use 2-MeBHPs as unambiguous biomarkers for any particular taxonomic group. This leaves open the question of whether 2-MeBHPs instead reflect a deeper underlying physiological function for the organisms that produce them. Notably, 43% of organisms with *hpnP* and 63% of *hpnP* sequences from metagenomes are plant associated, defined as bacteria that form commensal or mutualistic symbiotic interactions with land plants. Notable examples include

Bradyrhizobium spp., *Methylobacterium* spp. and *Nostoc* spp. (Figure 1); (Knani *et al.,* 1994; Bravo *et al.,* 2001; Meeks, 2009). To assess whether this observation was non-random, we used the hypergeometric test calculated with the number of plant-associated bacteria estimated at the species level to alleviate bias in the data set. We also performed this test with *shc*-containing bacteria and metagenomes. In all cases, we found the enrichment between (meta)genomes containing *hpnP* or *shc* and those that are plant associated to be significant (P-values $< 10^{-6}$): 46% and 51% of *hpnP*-containing bacteria and metagenomes are plant associated, as well as 24 and 30% of those containing *shc*, in contrast to only 9% of bacteria (Supplementary Table 2). There is thus a preferential association of 2-MeBHP production, and to a lesser extent hopanoid production, with plant symbiosis.

To expand on this, we investigated the possibility that metabolisms used to establish plant–microbe interaction might be correlated to *hpnP*. It is common for plant symbionts to provide fixed nitrogen to plants or utilize methanol, a byproduct of plant metabolism, as a carbon source (Bravo *et al.,* 2001; Gourion *et al.,* 2006; Meeks, 2009). We found that 74% of *hpnP*-containing bacteria had either *nifD* or *moxF*, which encode proteins necessary for nitrogen fixation and methanol utilization, respectively (Figure 1, middle ring). Using the hypergeometric test, we found *hpnP* to be significantly overrepresented among bacteria containing *nifD* or *moxF* (P-values $< 10^{-6}$; Supplementary Table 3). As the presence of *nifD* or *moxF* is not exclusive to plant symbionts, these numbers are likely overestimates, whereas the percentage of *hpnP*-containing bacteria that are plant associated is an underestimate of the number of *hpnP*-containing bacteria that are plant associated because not all organisms have been tested for the ability to form symbioses.

Consistent with this finding, the *hpnP* gene in *R. palustris* TIE-1 is regulated by an extracytoplasmic function transcription factor conserved in alphaproteobacteria that has a role in establishing plant symbioses in *Bradyrhizobium japonicum* and *Methylobacterium extorquens* (Gourion *et al.,* 2006, 2009). This factor induces *hpnP* expression in response to osmotic stress in *R. palustris* (Kulkarni *et al.,* 2013); osmolyte production by plants in the rhizosphere is well documented, and in some cases has been shown to promote plant–microbe symbioses (Miller and Wood, 1996; Khamar *et al.,* 2010).

Geobiological implications

Although the presence of *hpnP* significantly correlates with plant-associated bacteria in modern environments, the 2-methylhopane record predate the rise of land plants (Clarke *et al.,* 2011). Thus, ancient symbioses clearly cannot explain the presence of 2-methylhopanes in the remote past. A possible explanation for why today we find the capacity for 2-MeBHP production enriched in habitats containing microbe–plant associations is that the capacity to make 2-MeBHPs is selected by particular environmental conditions present in these habitats that are similar to those in the ancient depositional context. Specifically, we note that many modern environments containing 2-MeBHP producers comprise sessile microbial communities that have suboxia or anoxia, high osmolarity and limited fixed nitrogen; these same parameters have also been used to describe the depositional context of 2-methylhopanes. For example, increased 2-methylhopane indices have been measured in ancient sedimentary rocks recording ocean anoxic events, which may have favored nitrogen fixing organisms (Knoll *et al.,* 2007) and in sessile microbial mat and stromatolites, which contain high osmolytes in the form of extracellular polysaccharides and excreted small molecules (Summons *et al.,* 1999). Although none of the described niche parameters are solely responsible for the presence of *hpnP* or 2-MeBHPs, and 2-MeBHPs are not required for occupancy of this niche, their combination appears to be correlated with the capacity for 2-MeBHP production. Determining the underlying cellular role for 2-MeBHPs given the described niche is necessary to provide a more definitive interpretation for ancient 2-methylhopanes. In conclusion, our ecological data demonstrate that 2-MeBHPs cannot be used as taxonomic biomarkers for any particular group but suggest 2-MeBHPs may be diagnostic for the confluence of particular environmental parameters.

Acknowledgements

We acknowledge members of the Newman lab for constructive comments on the manuscript. This work was supported by grants from the Howard Hughes Medical Institute to DKN and a NASA award (NNX12AD93G) to DKN, ALS and RES. Research access to Yellowstone hot springs was granted to JRS from the Yellowstone Center for Resources. We thank V Orphan, D Des Marais, the NASA Ames Research Center and ESSA Exportadora del Sal, SA, de CV, Guerrero Negro, Baja, California Sur, Mexico for samples of Guerrero Negro microbial mats. We are grateful to E Allen, J Valliere, D Caron and A Lie for help with sample collection. JNR was supported by an NSF graduate fellowship, MLC by an Agouron Institute postdoctoral fellowship and PVW by a NASA Astrobiology Institute postdoctoral fellowship. DKN is an HHMI Investigator.

References

Biers EJ, Sun S, Howard EC. (2009). Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**: 2221–2229.

Brasier M, Green O, Lindsay J, Steele A. (2004). Earth's oldest (approximately 3.5 Ga) fossils and the 'early

Eden hypothesis': questioning the evidence. *Orig Life Evol Biosph* **34**: 257–269.
Brasier M, Green OR, Jephcoat AP, Kleppe AK, Van Kranendonk MJ, Lindsay JF *et al.* (2002). Questioning the evidence for Earth's oldest fossils. *Nature* **416**: 76–81.
Bravo J, Perzl M, Härtner T, Kannenberg EL, Rohmer M. (2001). Novel methylated triterpenoids of the gammacerane series from the nitrogen-fixing bacterium Bradyrhizobium japonicum USDA 110. *Eur J Biochem* **268**: 1323–1331.
Brocks JJ, Logan GA, Buick R, Summons RE. (1999). Archean molecular fossils and the early rise of eukaryotes. *Science* **285**: 1033–1036.
Brocks JJ, Pearson A. (2005). Building the biomarker tree of life. *Rev Mineral Geochem* **59**: 233–258.
Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.
Clarke JT, Warnock RCM, Donoghue PCJ. (2011). Establishing a time-scale for plant evolution. *New Phytol* **192**: 266–301.
Cooke MP, Talbot HM, Farrimond P. (2008). Bacterial populations recorded in bacteriohopanepolyol distributions in soils from Northern England. *Org Geochem* **39**: 1347–1358.
Dupraz C, Visscher PT. (2005). Microbial lithification in marine stromatolites and hypersaline mats. *Trends Microbiol* **13**: 429–438.
Etherington GJ, Dicks J, Roberts IN. (2005). Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics* **21**: 278–281.
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.
Garby TJ, Walter MR, Larkum AW, Neilan BA. (2012). Diversity of cyanobacterial biomarker genes from the stromatolites of Shark Bay, Western Australia. *Environ Microbiol* **15**: 1464–1475.
Gourion B, Rossignol M, Vorholt JA. (2006). A proteomic study of Methylobacterium extorquens reveals a response regulator essential for epiphytic growth. *Proc Natl Acad Sci USA* **103**: 13186–13191.
Gourion B, Sulser S, Frunzke J, Francez-Charlot A, Stiefel P, Pessi G *et al.* (2009). The PhyR-sigma(EcfG) signalling cascade is involved in stress response and symbiotic efficiency in Bradyrhizobium japonicum. *Mol Microbiol* **73**: 291–305.
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.
Hoehler TM, Bebout BM, Des Marais DJ. (2001). The role of microbial mats in the production of reduced gases on the early Earth. *Nature* **412**: 324–327.
Huang Y, Niu B, Gao Y, Fu L, Li W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**: 680–682.
Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.
Katoh K, Toh H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**: 286–298.
Khamar HJ, Breathwaite EK, Prasse CE, Fraley ER, Secor CR, Chibane FL *et al.* (2010). Multiple roles of soluble sugars in the establishment of Gunnera-Nostoc endosymbiosis. *Plant Physiol* **154**: 1381–1389.
Knani M, Corpe WA, Rohmer M. (1994). Bacterial hopanoids from pink-pigmented facultative methylotrophs (PPFMs) and from green plant surfaces. *Microbiology* **140**: 2755–2759.
Knoll AH, Summons RE, Waldbauer JR, Zumberge JE. (2007). The geological succession of primary producers in the oceans. In *The Evolution of Primary Producers in the SeaFalkowski P, Knoll AH (ed)Academic Press: Boston, pp 133–164.*
Kulkarni G, Wu C-H, Newman DK. (2013). The general stress response factor EcfG regulates expression of the C-2 hopanoid methylase HpnP in Rhodopseudomonas palustris TIE-1. *J Bacteriol* **195**: 2490–2498.
Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**: 127–128.
Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y *et al.* (2011). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40**: D123–D129.
Meeks JC. (2009). Physiological adaptations in nitrogen-fixing nostoc–plant symbiotic associations. In *Microbiology Monographs: Prokaryotic Symbionts in PlantsPawlowski K (ed)Vol. 8. Springer-Verlag: Munster, Germany, pp 181–205.*
Miller KJ, Wood JM. (1996). Osmoadaptation by rhizosphere bacteria. *Annu Rev Microbiol* **50**: 101–136.
Mulkidjanian AY, Koonin E V, Makarova KS, Mekhedov SL, Sorokin A, Wolf YI *et al.* (2006). The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci USA* **103**: 13126–13131.
Osburn MR, Sessions AL, Pepe-Ranney C, Spear JR. (2011). *Hydrogen-isotopic variability in fatty acids from Yellowstone National Park hot spring microbial communities. Geochim Cosmochim Acta* **75**: 4830–4845.
Ourisson G, Rohmer M, Poralla K. (1987). Prokaryotic hopanoids and other polyterpenoid sterol surrogates. *Annu Rev Microbiol* **41**: 301–333.
Pearson A, Flood Page SR, Jorgenson TL, Fischer WW, Higgins MB. (2007). Novel hopanoid cyclases from the environment. *Environ Microbiol* **9**: 2175–2188.
Pearson A, Leavitt WD, Sáenz JP, Summons RE, MC-M Tam, Close HG. (2009). Diversity of hopanoids and squalene-hopene cyclases across a tropical land-sea gradient. *Environ Microbiol* **11**: 1208–1223.
Pearson A, Rusch DB. (2009). Distribution of microbial terpenoid lipid cyclases in the global ocean metagenome. *ISME J* **3**: 352–363.
Rashby SE, Sessions AL, Summons RE, Newman DK. (2007). Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc Natl Acad Sci USA* **104**: 15099–15104.
Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR. (2008). Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455**: 1101–1104.
Rohmer M. (2010). Handbook of hydrocarbon and lipid microbiology. In *Chemistry and Physics of LipidsTimmis KN (ed).Springer Berlin Heidelberg: Berlin, Heidelberg.*

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

Sáenz JP, Eglinton TI, Summons RE. (2011). Abundance and structural diversity of bacteriohopanepolyols in suspended particulate matter along a river to ocean transect. *Org Geochem* **42**: 774–780.

Schopf JW, Packer B. (1987). Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science* **237**: 70–73.

Sessions AL, Zhang L, Welander PV, Doughty D, Summons RE, Newman DK. (2013). Identification and quantification of polyfunctionalized hopanoids by high temperature gas chromatography–mass spectrometry. *Org Geochem* **56**: 120–130.

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

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–D551.

Talavera G, Castresana J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564–577.

Talbot HM, Farrimond P. (2007). Bacterial populations recorded in diverse sedimentary biohopanoid distributions. *Org Geochem* **38**: 1212–1225.

Talbot HM, Rohmer M, Farrimond P. (2007a). Rapid structural elucidation of composite bacterial hopanoids by atmospheric pressure chemical ionization liquid chromatography/ion trap mass spectrometry. *Rapid Commun Mass Spectrometry* **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. (2003). 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 RE, Jahnke L, Cockell CS, Rohmer M, Farrimond P. (2008). Cyanobacterial bacteriohopanepolyol signatures from cultures and natural environmental settings. *Org Geochem* **39**: 232–263.

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

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.

Walter MR, Grotzinger JP, Schopf JW. (1992). Proterozoic stromatolites. In *The Proterozoic Biosphere: A Multidisciplinary StudySchopf JW, Klien C (eds).Cambridge University Press: Cambridge, pp 253–260.*

Welander PV, Coleman ML, Sessions AL, Summons RE, Newman DK. (2010). Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proc Natl Acad Sci USA* **107**: 8537–8542.

Welander PV, Doughty DM, Wu C-H, Mehay S, Summons RE, Newman DK. (2012). Identification and characterization of Rhodopseudomonas palustris TIE-1 hopanoid biosynthesis mutants. *Geobiology* **10**: 163–177.

Xu Y, Cooke MP, Talbot HM, Simpson MJ. (2009). Bacteriohopanepolyol signatures of bacterial populations in Western Canadian soils. *Org Geochem* **40**: 79–86.

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