Synthesis of arborane triterpenols by a bacterial oxidosqualene cyclase

Amy B. Banta[a,1], Jeremy H. Wei[a,1], Clare C. C. Gill[a], José-Luis Giner[b], and Paula V. Welander[a,2]

[a]Department of Earth System Science, Stanford University, Stanford, CA 94305; and [b]Department of Chemistry, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210

Edited by John M. Hayes, Woods Hole Oceanographic Institution, Berkeley, CA, and approved November 16, 2016 (received for review October 17, 2016)

Cyclic triterpenoids are a broad class of polycyclic lipids produced by bacteria and eukaryotes. They are biologically relevant for their roles in cellular physiology, including membrane structure and function, and biochemically relevant for their exquisite enzymatic cyclization mechanism. Cyclic triterpenoids are also geobiologically significant as they are readily preserved in sediments and are used as biomarkers for ancient life throughout Earth's history. Isoarborinol is one such triterpenoid whose only known biological sources are certain angiosperms and whose diagenetic derivatives (arboranes) are often used as indicators of terrestrial input into aquatic environments. However, the occurrence of arborane biomarkers in Permian and Triassic sediments, which predates the accepted origin of angiosperms, suggests that microbial sources of these lipids may also exist. In this study, we identify two isoarborinol-like lipids, eudoraenol and adriaticol, produced by the aerobic marine heterotrophic bacterium *Eudoraea adriatica*. Phylogenetic analysis demonstrates that the *E. adriatica* eudoraenol synthase is an oxidosqualene cyclase homologous to bacterial lanosterol synthases and distinct from plant triterpenoid synthases. Using an *Escherichia coli* heterologous sterol expression system, we demonstrate that substitution of four amino acid residues in a bacterial lanosterol synthase enabled synthesis of pentacyclic arborinols in addition to tetracyclic sterols. This variant provides valuable mechanistic insight into triterpenoid synthesis and reveals diagnostic amino acid residues to differentiate between sterol and arborinol synthases in genomic and metagenomic datasets. Our data suggest that there may be additional bacterial arborinol producers in marine and freshwater environments that could expand our understanding of these geologically informative lipids.

triterpene synthase | isoarborinol | sterol | biomarker | natural products

Cyclic triterpenoids are a broad class of lipids produced by diverse bacteria and eukaryotes (1). The most studied of these molecules are the tetracyclic sterols (e.g., cholesterol) and their derivatives (e.g., steroid hormones), which are essential in eukaryotes and play critical roles in membrane structure and in cellular signaling (2–4). In addition to sterols, plants synthesize a diverse array of cyclic triterpenoids that have a variety of functions, including defense against pests and pathogens (5–7). A few bacteria have been shown to produce sterols (8), however, the most common bacterial cyclic triterpenoids are the pentacyclic hopanoids, which are thought to function as "sterol surrogates" in bacterial membranes (9, 10). Although the majority of interest in cyclic triterpenoids stems from their essential physiological roles and unique enzymatic biosynthesis (5, 7, 11), these lipids are also significant from a geological perspective. Cyclic triterpenoids are quite recalcitrant and, as a result, are well preserved in sedimentary rocks and can serve as geological biomarkers that link organisms to environments deep in Earth's history (12).

The interpretation of geological biomarkers is primarily based on the occurrence of their diagenetic precursors in extant organisms and/or their prevalence in specific ecosystems (12). However, incomplete understanding of the distribution and function of potential biomarkers in modern systems can lead to inconsistencies in their interpretations. For example, arborane biomarkers are thought to be derived from isoarborinol, an unusual pentacyclic triterpenol whose only known extant sources are certain flowering plants (13–16). Thus, arborane biosignatures are considered robust indicators of angiosperms and of terrestrial input into marine and lacustrine environments. However, the detection of arborane signatures in Permian and Triassic sediments (17–19), which predates the accepted first appearance of angiosperms, as well as compound-specific ^{13}C values that are inconsistent with plant sources, led researchers to propose that there were microbial sources of isoarborinol (19–21). These sources, however, remain undiscovered.

The discovery of arborinol lipids in a microbe would also be significant from a biochemical perspective. Cyclic triterpenoid lipids are synthesized by cyclization of a 30-carbon acyclic isoprenoid substrate through a series of carbocation intermediates in the central cavity of a terpene cyclase (class II) enzyme (11, 22). These enzymes can be distinguished based on three general characteristics: the use of squalene or oxidosqualene as the initial substrate, the conformation of the acyclic substrate in the more energetically favorable all-chair (CCC) versus the more strained chair−boat−chair (CBC or B boat), and the total number and size of the rings generated in the final product (e.g., tetracyclic sterols versus pentacyclic hopanoids and isoarborinol) (5, 7, 23). Hopanoid synthases fold squalene into the CCC conformation and generate a pentacyclic structure (11, 24), whereas sterol synthases cyclize oxidosqualene in the CBC conformation and generate a tetracyclic structure (25). An arborinol synthase represents a unique combination of these characteristics. It is similar to a sterol cyclase in that it cyclizes oxidosqualene in the CBC conformation (6, 7, 23) but differs in that its final product

Significance

Polycyclic lipids produced by bacteria and eukaryotes can be preserved in sedimentary rocks for millions of years. These ancient lipids can function as "molecular fossils" or biomarkers that can inform us about the types of organisms and environments on early Earth. However, proper interpretation of these biomarkers requires a comprehensive understanding of the taxonomic distribution, biosynthesis, and physiological function of these lipids in modern organisms. In this study, we discover that a marine bacterium produces two arborinols, a class of lipids previously identified only in flowering plants. This discovery addresses a current incongruity in biomarker signatures and also provides insight into the evolution of the biosynthetic pathways of biomarker lipids.

Author contributions: A.B.B., J.H.W., and P.V.W. designed research; A.B.B., J.H.W., J.-L.G., and C.C.C.G. performed research; A.B.B., J.H.W., J.-L.G., and P.V.W. analyzed data; and A.B.B., J.H.W., J.-L.G., and P.V.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.B.B. and J.H.W. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: welander@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617231114/-/DCSupplemental.

PNAS PNAS PNAS

has the pentacyclic structure similar to a bacterial hopanoid (23). Various hopanoid, sterol, and plant triterpenoid synthases have been characterized, including a rice isoarborinol synthase (IAS) (6). However, a microbial enzyme that cyclizes oxidosqualene in the CBC conformation to an arborinol lipid has yet to be identified.

Here, we present two pentacyclic triterpenols, eudoraenol and adriaticol, produced by the marine heterotrophic bacterium *Eudoraea adriatica* (26). These molecules represent the first arborane lipids identified outside of the plant kingdom. Further, the *E. adriatica* eudoraenol synthase (EUS) is the first example of a microbial enzyme that cyclizes an oxidosqualene precursor in the CBC conformation to a pentacyclic triterpenol. EUS is phylogenetically distinct from plant triterpene synthases, indicating that it is not derived from the known eukaryotic IAS. To further understand the relationship between these enzymes, we created a bacterial lanosterol synthase (LAS) gain of function variant with substitutions in four key amino acid residues that synthesizes both tetracyclic and pentacyclic triterpenols. These results together support the hypotheses that synthesis of the arborane skeleton has likely arisen more than once within the oxidosqualene cyclase (OSC) family and that Permian and Triassic arborane biomarkers are likely from a microbial source.

Results

E. adriatica Produces Unique Triterpenols of the Arborinol Class. An OSC homolog was previously identified in the aerobic heterotrophic bacterium *E. adriatica* (8, 27), a member of the family *Flavobacteriaceae* in the phylum *Bacteriodetes* that was isolated from surface waters of the Adriatic Sea (26). Our lipid analysis of *E. adriatica* identified a trace amount of lanosterol along with two potential sterol-like triterpenols whose spectra were distinct from any that had been previously published (Fig. 1 and Fig. S1). The two compounds, found in a 7:1 ratio, were purified by reversed-phase HPLC (RP-HPLC), and their structures were determined using 800-MHz ^{1}H NMR (Fig. S2 and Tables S1 and S2). The spectra of both compounds showed six methyl singlets and two methyl doublets typical of the hopane skeleton. Heteronuclear multiple bond correlation (HMBC) spectra localized the double bonds at positions 12 and 8, indicating that these were isomers of the fernanes neomotiol and isomotiol, respectively (Fig. S2). Mass spectra confirmed the positions of the double bonds with a diagnostic ion at m/z = 218 for the major Δ^{12} compound and its derivatives, and diagnostic ions at m/z = 259, 301, and 331 for the free Δ^{8} compound, its acetate, and its trimethylsilyl (TMS) ether, respectively (Fig. S2) (28). The stereochemical configurations were determined by chemical correlation with isoarborinol (16) and boehmerol (29), both of which arise from CBC cyclization (7), using acid-catalyzed isomerization. The major triterpenol isomerized to boehmerol and the minor component was formed in the acid-catalyzed isomerization of isoarborinol (Fig. S2). These data demonstrate that *E. adriatica* is producing two triterpenoids of the arborane class that we have named eudoraenol and adriaticol, respectively. These are the first new OSC products with a constitutional hopanoid skeleton to be discovered since boehmerol was reported 30 y ago (29), as well as the first evidence of bacterial synthesis of pentacyclic CBC triterpenols derived from oxidosqualene.

E. adriatica and Plant OSC Are Phylogenetically Distinct. Because the phylogenetic distribution of OSC homologs in bacteria is sporadic, it was unclear whether the *E. adriatica* OSC was derived by evolutionary diversification of a bacterial or eukaryotic OSC or by acquisition of a plant IAS through horizontal gene transfer. To address this, we determined the phylogenetic relationship of over 800 terpenoid cyclase homologs obtained from the Joint Genome Institute (JGI) genomic and metagenomic databases using maximum likelihood analysis (30, 31). We found that the *E. adriatica* cyclase does not cluster with plant triterpenoid synthases and, in particular, is distinct from the rice IAS (Fig. 2*A*). Instead, the *E. adriatica* OSC branches within a distinct clade of OSCs from three single-cell genomes of *Eudoraea* species isolated from the North Sea as well as 16 metagenomic OSC sequences (Fig. 2*B*). The metagenomic sequences separate into two distinct clades reflecting different ecosystems—one from marine sources and the other from lacustrine sources. These data indicate that eudoraenol and adriaticol may be produced in freshwater as well as marine environments and that there could be additional extant bacterial producers of similar triterpenoids.

E. adriatica OSC Synthesizes Eudoraenol and Adriaticol Directly from Oxidosqualene. Detection of a minor amount of lanosterol in *E. adriatica* extracts raised the possibility that the *E. adriatica* OSC does not synthesize eudoraenol and adriaticol directly. Rather, it was conceivable that the *E. adriatica* OSC first synthesizes a partially cyclized CBC compound, and then additional protein(s) subsequently modify this product to generate eudoraenol and adriaticol. Precedence for this multiple-enzyme scenario exists in the synthesis of the pentacyclic triterpenoid tetrahymanol (32). In this case, eukaryotes use a single cyclase to synthesize tetrahymanol directly from squalene, but bacteria use a two enzyme system with the first cyclizing squalene to diploptene and the second catalyzing a ring expansion to tetrahymanol.

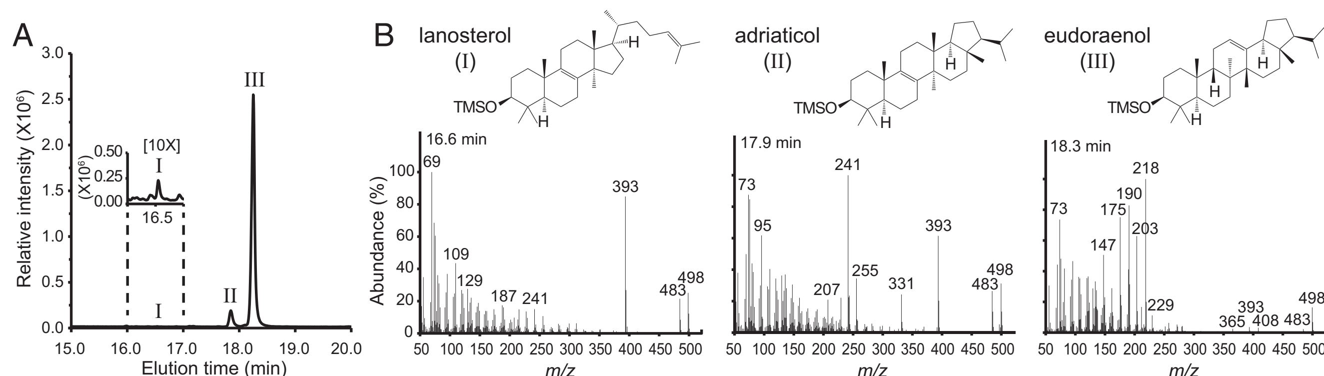
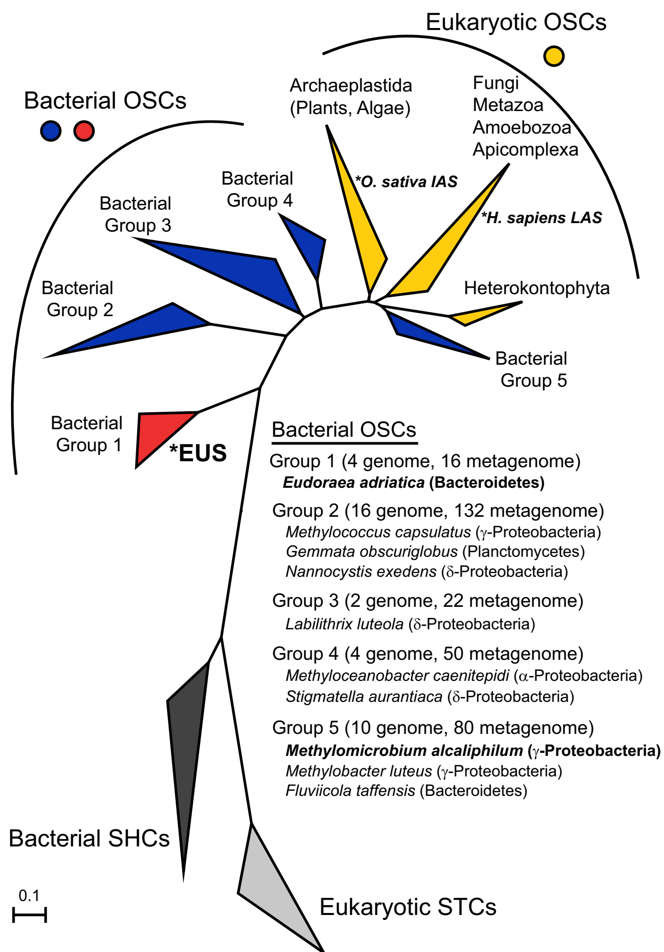


Fig. 1. Polycyclic triterpenols detected in *E. adriatica*. (*A*) GC-MS total ion chromatogram (TIC) of the alcohol-soluble fraction of a total lipid extract (TLE), derivatized to TMS ethers, of *E. adriatica* showing three distinct peaks eluting at 16.6 min (I), 17.9 min (II) and 18.3 min (III). (*Inset*) Peak I visible when a 10× concentrate was loaded. Triterpenols shown here constitute ∼1% of the TLE. (*B*) Mass spectrum (MS) of compound I identified as lanosterol by comparison with published spectra. MS of compound II with structure determined by NMR and designated adriaticol. MS of compound III with structure determined by NMR and designated eudoraenol. NMR data are listed in Table S1 and shown in Fig. S2. GC-MS analysis of acetate esters is shown in Fig. S1.



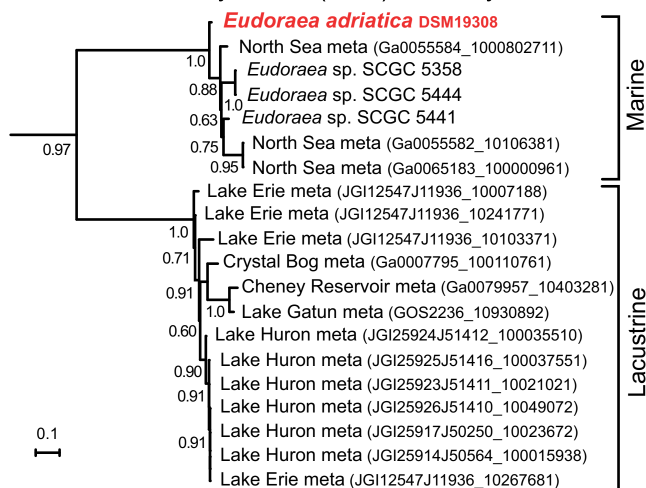


Fig. 2. Phylogenetic analysis of EUS. (*A*) Unrooted maximum likelihood phylogenetic tree of OSC homologs identified in genomes and metagenomes (851 sequences) with bacterial squalene-hopene cyclases (SHC) and eukaryotic squalene tetrahymanol cyclases (STC) as the outgroup. Sequences in each branch have been collapsed for clarity. Bacterial clades are colored in blue or red, with the total number of genome and metagenome sequences as well as representative cultured organisms listed for each group. (*B*) Expanded Bacterial OSC Group 1 branch from the phylogenetic tree in *A*. JGI locus tag numbers are listed for each metagenome sequence in parentheses. The scale bar indicates 0.1 changes per nucleotide site.

To determine if the *E. adriatica* OSC cyclizes oxidosqualene directly to eudoraenol and adriaticol, we developed an inducible heterologous expression system using three plasmids in *Escherichia coli.* The first plasmid increases overall isoprenoid synthesis by overexpression of the mevalonate pathway (33). The second plasmid enables synthesis of the acyclic triterpenoids squalene and/or oxidosqualene by encoding the squalene synthase (*sqs*) gene alone or together with the squalene epoxidase (*smo*) gene from the sterol-producing bacterium *Methylomicrobium alcaliphilum*. Finally, the third plasmid encodes putative OSC (*osc*) genes. Expression of *M. alcaliphilum osc*, which encodes an LAS, in an oxidosqualene-producing *E. coli* strain resulted in lanosterol synthesis (Fig. 3*A* and Fig. S3) (32). Expression of *E. adriatica osc* in an oxidosqualene-producing *E. coli* strain resulted in synthesis of eudoraenol and adriaticol as well as a trace amount of lanosterol, as observed in *E. adriatica* lipid extracts (Fig. 3*B* and Fig. S3). Expression of the *E. adriatica osc* in a squalene-producing *E. coli* strain did not result in any cyclic triterpenoid production (Fig. S3). These results confirm that *E. adriatica* OSC synthesizes the pentacyclic triterpenols eudoraenol and adriaticol directly from an oxidosqualene precursor.

Identification of Key Residues Distinguishing Pentacyclic Versus Tetracyclic Triterpenol Synthases. The predominant synthesis of pentacyclic arborinol lipids rather than tetracyclic sterols by *E. adriatica* eudoraenol synthase indicates that homology alone is not sufficient to predict the full lipid profile of a putative cyclase. We took a comparative analysis approach to determine specific amino acid residues that are necessary for synthesis of the fifth (E) ring structure that could aid in identification of other OSCs that potentially synthesize pentacyclic lipids. First, we identified amino acid residues that are conserved in EUS and conserved among sterol synthases but that differ between the groups by aligning *E. adriatica* EUS with a diversity of bacterial OSCs known to synthesize tetracyclic sterols (Fig. 4*A*) (8). We further selected residues that are likely to be in the active site cavity near the site of formation of the E ring by alignment with the *Homo sapiens* LAS X-ray crystal structure [Protein Data Bank (PDB) ID code 1W6K] (Fig. 4*B*) (34). We then made reciprocal substitutions by changing the identity of residues in *M. alcaliphilum* LAS to those in *E. adriatica* EUS, and vice versa, and determined the lipid profile of these variant enzymes using our *E. coli* heterologous expression system.

Two positions that are highly conserved in LAS, H232 (histidine) and Y503 (tyrosine) (*H. sapiens* numbering), had a substantial difference in identity from the homologous residues in *E. adriatica* EUS [Y164 (tyrosine) and V428 (valine)] (Fig. 4*A*). Previous studies of the *Saccharomyces cerevisiae* LAS indicated that this hydrogen-bonded H–Y pair plays a key role in both cyclization and terminal proton abstraction to yield the tetracyclic structure of lanosterol (35–38), suggesting that these residues could potentially contribute to the formation of the eudoraenol and adriaticol pentacyclic structure. To test this, we changed these residues, both individually and in combination, in *M. alcaliphilum* LAS to the corresponding *E. adriatica* EUS residues (Table S3; lipid structures shown in Fig. S4). The LAS H254Y single variant had significantly reduced oxidosqualene cyclization in general, whereas the Y521V variant synthesized parkeol, a lanosterol isomer, in addition to lanosterol (Table S3). Although we found that the H254Y/Y521V double variant synthesized a trace amount of adriaticol in addition to lanosterol and parkeol, those two changes alone were not sufficient to enable LAS to synthesize the main EUS pentacyclic structure, eudoraenol.

Continuing our comparative analysis, we found that *E. adriatica* EUS residues S162 (serine) and Y618 (tyrosine) not only differ from the corresponding *M. alcaliphilum* LAS residues W252 (tryptophan) and N717 (asparagine) but are also adjacent to the above residues (H254 and Y521) in the active site cavity (Fig. 4*B*). Although the identities of these four residues are not completely

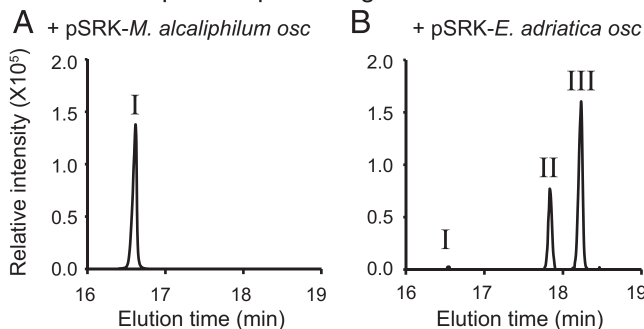


Fig. 3. Sterol and eudoraenol synthesis in an *E. coli* heterologous OSC expression system. GC-MS extracted ion chromatograms (EIC *m/z* 498) of the alcohol-soluble fraction of a TLE (derivatized to TMS ethers) of an oxidosqualene-producing strain of *E. coli* overexpressing either (*A*) *M. alcaliphilum osc* or (*B*) *E. adriatica osc* from plasmid pSRK. Lipid content of peaks was identified by MS as follows: 16.6 min lanosterol (I), 17.8 min adriaticol (II), and 18.2 min eudoraenol (III). TICs of samples and controls are shown in Fig. S3.

conserved in all LAS homologs (24), they tend to covary and always differ from those in EUS. We changed these residues alone and in combination with the previous substitutions in the *M. alcaliphilum* LAS and tested the variant proteins to determine whether these residues could contribute to the synthesis of the pentacyclic arborinols. The W252S/N717Y double variant synthesized the pentacyclic lipids adriaticol and isoarborinol in addition to tetracyclic parkeol and lanosterol. However, the single variants (W252S or N717Y alone) only synthesized parkeol and lanosterol (Table S3), indicating that these residues together affect synthesis of the E ring. An *M. alcaliphilum* LAS variant with all four substitutions (W252S, H254Y, Y521V, and N717Y) synthesized tetracyclic lanosterol and parkeol and pentacyclic adriaticol and isoarborinol as well as trace amounts of pentacyclic isomers eudoraenol and boehmerol, with all products retaining the CBC conformation (Fig. 4*C*, Fig. S5, and Table S3).

Finally, we constructed the reciprocal substitutions of these four residues in *E. adriatica* EUS to the identity of those in *M. alcaliphilum* LAS. Although three of the four single substitution variants still synthesized pentacyclic structures, the identity of those structures shifted with each substitution (Table S3). The EUS Y164H single variant synthesized eudoraenol but not adriaticol, whereas the EUS V428Y substitution resulted in isoarborinol synthesis in addition to eudoraenol and adriaticol. The EUS Y618N also synthesized both eudoraenol and adriaticol, but the dominant product was a tetracyclic structure, which we have tentatively identified as protosta-20 (22), 24-dien-3-ol. The S162W substitution completely eliminated both tetracyclic and pentacyclic lipid synthesis, which made it difficult to interpret the results of subsequent combined substitutions (Table S3). Nonetheless, these single substitution results suggest that the collective identity of these four amino acid residues is important not only for the synthesis of the additional ring in pentacyclic versus tetracyclic triterpenoids but also for the positioning of the double bonds and the stereochemistry of the methyl groups in the final product.

Discussion

Bioinformatics coupled to lipid analyses and protein characterization have contributed significantly to our understanding of the taxonomic distribution of microbial lipid biomarkers, the enzymatic mechanisms of their synthesis, and the potential evolutionary relationships of their biosynthetic pathways (39). Here, this combined approach revealed two triterpenols that not only address a current

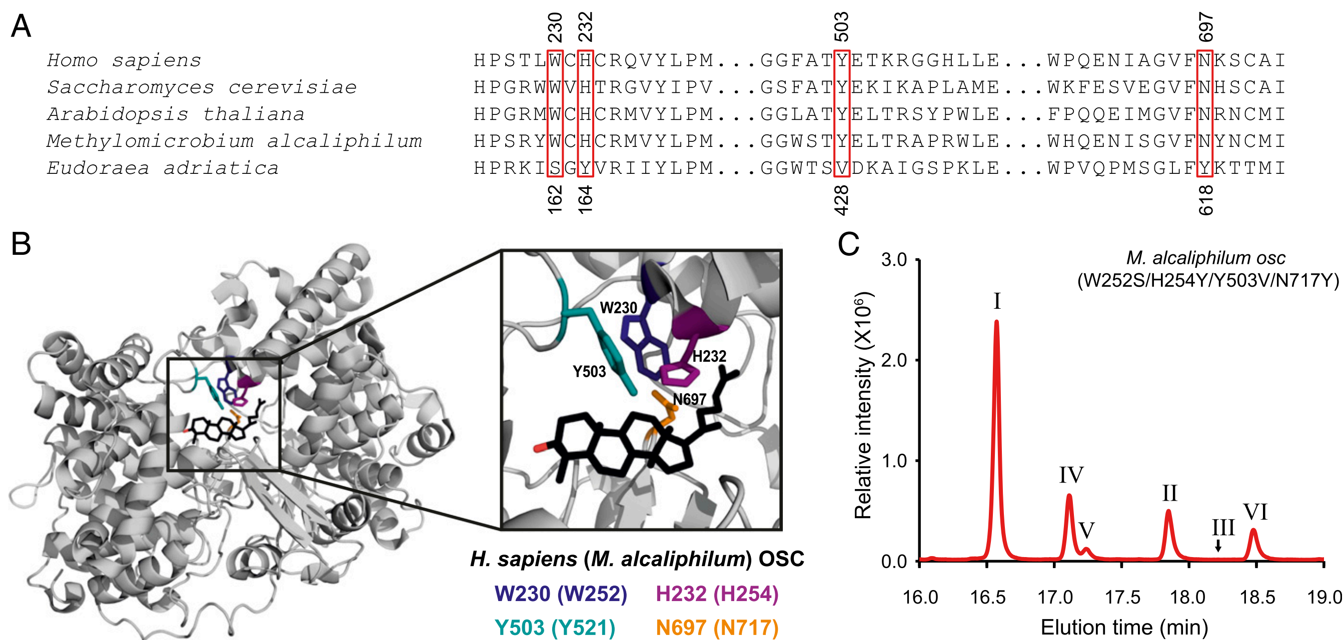


Fig. 4. Identification of key residues necessary for synthesis of pentacyclic arborinol versus tetracyclic sterols. (*A*) Partial amino acid sequence alignment of selected OSCs. Top numbers reflect positions in *H. sapiens* and bottom numbers reflect those in *E. adriatica*. Red boxes indicate positions that were changed by site-directed mutagenesis. (*B*) X-ray crystal structure of *H. sapiens* OSC (gray cartoon representation; PDB ID code 1W6K) bound to lanosterol (black stick representation with the C3–OH in red). Side chains of amino acids of interest are shown in stick representation in color as indicated (*H. sapiens/M. alcaliphilum* numbering). (*C*) GC-MS TIC of the alcohol-soluble fraction of a TLE (derivatized to TMS ethers) of *E. coli* oxidosqualene production strain overexpressing the *M. alcaliphilum* OSC W252S/H254Y/Y503V/N717Y variant, demonstrating partial conversion of an LAS to an EUS. Peaks labeled with Roman numerals have been identified by MS and/or NMR as follows: lanosterol (I), adriaticol (II), eudoraenol (III), parkeol (IV), boehmerol (V), and isoarborinol (VI). Structures of these lipids are shown in Fig. S4, and mass spectra are shown in Fig. S5.

incongruity in biomarker signatures but also provide mechanistic insight into polycyclic triterpenoid biosynthesis.

Eudoraenol and adriaticol are unique in both their chemical structure and biological source. Their structures are similar to isoarborinol, a C3-oxygenated pentacyclic triterpenol whose only known biological source is certain families of angiosperms (13). However, some have hypothesized that there must be microbial sources for arborinol lipids because their diagenetic derivatives, arboranes, have been identified in geologic samples whose deposition is inconsistent with the distribution of isoarborinol in modern organisms (17–21). Although *E. adriatica* does not produce isoarborinol per se, it does synthesize an isoarborinol isomer, adriaticol. This represents the first bacterial source of this class of lipids, thereby linking arborinol/arborane compounds to a potential bacterial source through geologic time.

In addition, the *E. adriatica* EUS is the first example of a triterpenoid synthase from a bacterium that cyclizes oxidosqualene to a pentacyclic triterpenol, which is significant from both an evolutionary and biochemical perspective. Phylogenetic analysis of an isoarborinol synthase (IAS) from *Oryza sativa* (rice) demonstrated that it was recently derived from a plant cycloartenol synthase (6). Our analysis demonstrates that *E. adriatica* EUS is phylogenetically distinct from *O. sativa* IAS and most likely was not acquired via horizontal gene transfer from a plant source. Rather, this enzyme likely evolved separately within bacteria, being either derived from an LAS consistent with the proposed evolutionary scheme of Fischer and Pearson (23) or instead from a squalene-hopene cyclase consistent with the proposals of Ourisson et al. (21). The identification of EUS now enables structural and biochemical studies of this bacterial arborinol class of enzymes that could provide experimental evidence for these evolutionary scenarios.

Further, structural and enzymatic studies of EUS could provide insight into the cyclization mechanism of triterpene synthases—the key step in determining the basic cyclic structure of various polycyclic triterpenoids (35). Using a reverse genetics approach guided by amino acid variations between EUS and LAS, we demonstrated that substitution of four amino acid residues (W230/252S, H232/254Y, Y503/521V, and N697/717Y; *H. sapiens*/*M. alcaliphilum* numbering) in the LAS active site cavity resulted in a gain of function variant that could cyclize an additional (E) ring to synthesize pentacyclic as well as tetracyclic triterpenoids. We hypothesize that these four substitutions allow for nondiscriminate cyclization (40) by LAS resulting in increased synthesis of tetracyclic parkeol as well as synthesis of various pentacyclic triterpenols. Given the close proximity of these four residues, these substitutions may also alter the LAS cavity to accommodate a bulkier pentacyclic structure. Single amino acid substitutions of these homologous residues in yeast and fungal LAS impact various aspects of the cyclization reaction, including carbocation stabilization, backbone rearrangement, and deprotonation of the final product (11, 34, 38, 41–43). However, in those studies, substitutions only inhibit or alter synthesis of tetracyclic sterols and do not enable the synthesis of pentacyclic triterpenoids. A recent study of a plant triterpene synthase, the *Avena strigosa* (oat) β-amyrin synthase (SAD1), demonstrated that substitution of an amino acid residue adjacent to the active site cavity disrupted formation of the E ring of the pentacyclic triterpenol β-amyrin (44). This substitution, S725F (homologous to *H. sapiens* LAS residue S699), is two amino acid residues downstream of our *M. alcaliphilum* LAS N717Y substitution, which enabled the synthesis of pentacyclic triterpenoids. Even though the β-amyrin synthase differs from lanosterol and arborinol synthases in that it cyclizes oxidosqualene in the CCC rather than the CBC conformation (5, 7), these data demonstrate that this region is critical to the formation of the E ring in both plant and bacterial triterpenoid synthases. Thus, our studies together underscore the functional diversity of triterpenoid synthases and demonstrate the potential for engineering the specificity of these cyclases to synthesize novel triterpenoids and other secondary metabolites.

Finally, the four EUS amino acid residues identified in this study can now be used as diagnostic markers to discover other potential sources of arborinols and related lipids. Although the distribution of EUS in genomic databases is currently restricted to *Eudoraea* species (perhaps reflecting a sampling bias in genomic databases), analysis of metagenomic sequences reveals that there are potentially other marine (e.g., North Sea) or lacustrine (e.g., Lake Huron) bacterial sources. Thus, using the *E. coli* heterologous expression system developed here, we can now experimentally determine if any of these other putative arborinol synthases can synthesize eudoraenol, isoarborinol, or perhaps other triterpenoids. Ultimately, this combined bioinformatics and biochemical approach should enable us to identify additional unique biomarker lipid synthesis enzymes, analyze their products, and link them to cultured and/or uncultured organisms as well as ancient and/or modern environments. This, in turn, will inform our understanding of the evolutionary history of biomarker lipids and allow for more robust interpretations of their occurrence in the rock record.

Materials and Methods

Bacterial Culture. Strains used in this study are listed in Table S4. *E. adriatica* DSM 19308 was cultured in Bacto Marine Broth (Difco 2216) at 30 °C with shaking at 225 rpm. *E. coli* was cultured in lysogeny broth or terrific broth (TB) at 30 °C or 37 °C with shaking at 225 rpm. Media for *E. coli* was supplemented, if necessary, with gentamicin (15 μg/mL), carbenicillin (100 μg/mL), and/or chloramphenicol (20 μg/mL).

Molecular Cloning. Plasmids and oligonucleotides used in this study are listed in Tables S5 and S6. Details of molecular cloning techniques are described in *SI Materials and Methods* (45).

Heterologous Expression. Expression strains used in this study are listed in Table S7. *E. coli* DH10B strains harboring three plasmids, a pTrc99a derivative, a pSRKgm derivative, and pJBEI2997 (Addgene plasmid #35151) (33, 46, 47), were cultured at 37 °C, shaking in TB supplemented with chloramphenicol, carbenicillin, and gentamicin until midexponential phase. Expression was induced with 500 μM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 30 h to 40 h at 30 °C, shaking at 225 rpm.

Lipid Extraction and Analysis. Lipids were extracted from cells harvested from 4 mL to 50 mL of bacterial culture using a modified Bligh–Dyer extraction method (48, 49). Cells were first sonicated in 10:5:4 (v:v:v) methanol (MeOH): dichloromethane (DCM):water, and then the phases were separated by mixing with twice the volume of 1:1 (v:v) DCM:water followed by incubation at −20 °C and centrifugation. The organic phase was transferred to a new vial where solvents were evaporated under N_2 to yield the total lipid extract (TLE). The alcohol-soluble fractions of some TLEs were further purified by Si column chromatography (50). Lipids were derivatized to either acetate esters with 1:1 (v:v) pyridine:acetic anhydride or to TMS ethers with 1:1 (v:v) bis(trimethylsilyl)trifluoroacetamide (BSTFA):pyridine before analysis by gas chromatography–mass spectrometry (GC-MS). Further details of the lipid extraction and analysis techniques are described in *SI Materials and Methods*.

NMR Analysis. The TLEs were saponified and fractionated by preparative TLC as described in *SI Materials and Methods*. The triterpenol fraction was further fractionated by reversed-phase HPLC and characterized by NMR using a Bruker Avance III HD with an Ascend 800-MHz magnet and a 5-mm triple resonance inverse (TCI) cryoprobe at 30 °C using deuterated chloroform ($CDCl_3$) as the solvent. Calibration was by the residual solvent signal (7.26 ppm). The relative proportions of the triterpenols were determined by the integrals of the HPLC differential refractometer signal and the integrals of the ^{1}H NMR spectra. Acid-catalyzed isomerization of eudoraenol was carried out in dry $CDCl_3$ using 1% trifluoroacetic acid (TFA) with ^{1}H NMR monitoring as described in *SI Materials and Methods*. The major product was determined to be boehmerol, by comparison of its ^{1}H NMR spectrum with that of an authentic standard (51). Acid-catalyzed isomerization of isoarborinol [obtained from sorghum (52)] was carried out in dry $CDCl_3$ using 2% (vol/vol)

methanesulfonic acid as described in *SI Materials and Methods*. The products were determined to be adriaticol and an unknown triterpenol in a 2:1 ratio. The unknown triterpenol is likely to be the Δ^7 9α-isomer, in analogy to the product of isomerization of lanosterol (53).

Bioinformatics Analysis. We identified 941 triterpene cyclases, including squalene-hopene cyclases, squalene-tetrahymanol cyclases, and various OSCs using *Methylococcus capsulatus* Bath OSC (locus tag: MCA2873) to query the JGI Integrated Microbial Genomes & Microbiomes (IMG/M) databases (31) using the basic local alignment search tool for proteins (BLASTP) (54). For phylogenetic analysis of metagenome sequences, we selected only those that were larger than 400 amino acids, for a final total of 851 sequences. Protein sequences were aligned via Multiple Sequence Comparison by Log-Expectation (MUSCLE) (55) using Geneious (Biomatters Limited) and large gaps were removed from metagenomic sequence alignments using the Gblocks server (56). Maximum likelihood trees were constructed by maximum likelihood (PhyML) (30) using the LG+gamma model, four gamma rate categories, 10 random starting trees, nearest-neighbor interchanges (NNI) branch swapping, and substitution parameters estimated from the data. OSC trees were generated and edited through the interactive tree of life (iTOL) (57), using squalene-hopene cyclases and squalene-tetrahymanol cyclases as the outgroup.

ACKNOWLEDGMENTS. We thank members of the P.V.W. laboratory and Prof. Roger E. Summons for helpful discussions of this work. This study was supported by National Science Foundation Grants EAR-1451767 (to P.V.W.) and OCE-1061957 (to J.L.G.). The acquisition of the 800-MHz NMR spectrometer was made possible by National Institutes of Health Grant S10 OD012254.

1. Volkman JK (2005) Sterols and other triterpenoids: Source specificity and evolution of biosynthetic pathways. *Org Geochem* 36(2):139–159.
2. Parks LW, Casey WM (1995) Physiological implications of sterol biosynthesis in yeast. *Annu Rev Microbiol* 49:95–116.
3. Xu F, et al. (2005) Dual roles for cholesterol in mammalian cells. *Proc Natl Acad Sci USA* 102(41):14551–14556.
4. Brown MS, Goldstein JL (2009) Cholesterol feedback: From Schoenheimer's bottle to Scap's MELADL. *J Lipid Res* 50(Suppl):S15–S27.
5. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A (2014) Triterpene biosynthesis in plants. *Annu Rev Plant Biol* 65(65):225–257.
6. Xue Z, et al. (2012) Divergent evolution of oxidosqualene cyclases in plants. *New Phytol* 193(4):1022–1038.
7. Eschenmoser A, Arigoni D (2005) Revisited after 50 years: The stereochemical interpretation of the biogenetic isoprene rule for the triterpenes. *Helv Chim Acta* 88(12):3011–3050.
8. Wei JH, Yin X, Welander PV (2016) Sterol synthesis in diverse bacteria. *Front Microbiol* 7(990):990.
9. Kannenberg EL, Poralla K (1999) Hopanoid biosynthesis and function in bacteria. *Naturwissenschaften* 86(4):168–176.
10. Sáenz JP, et al. (2015) Hopanoids as functional analogues of cholesterol in bacterial membranes. *Proc Natl Acad Sci USA* 112(38):11971–11976.
11. Abe I (2007) Enzymatic synthesis of cyclic triterpenes. *Nat Prod Rep* 24(6):1311–1331.
12. Summons RS, Lincoln SA (2012) Biomarkers: Informative molecules for studies in geobiology. *Fundamentals of Geobiology*, eds Knoll AH, Canfield DE, Konhauser KO (Blackwell, West Sussex, UK), 1st Ed, pp 269–296.
13. Peters KE, Waters CC, Moldowan JM (2007) Biomarkers and Isotopes in Petroleum Exploration and Earth History, The Biomarker Guide (Cambridge Univ Press, Cambridge, UK), Vol 2.
14. Hemmers H, Gulz PG, Marner FJ, Wray V (1989) Pentacyclic triterpenoids in epicuticular waxes from *Euphorbia lathyris* L, *Euphorbiaceae*. *Z Naturforsch C* 44(3-4):193–201.
15. Ohmoto T, Nikaido T, Nakadai K, Tohyama E (1970) [Studies on the triterpenoids and the related compounds from gramineae plants. VII]. *Yakugaku Zasshi* 90(3):390–393.
16. Vorbrueggen H, Djerassi C, Pakrashi SC (1963) Arborinol, ein neuer triterpen-typus. *Ann Chem Justus Liebig* 668(1):57–76.
17. Hauke V, et al. (1992) Novel triterpene derived hydrocarbons of arborane fernane series in sediments: Part I. *Tetrahedron* 48(19):3915–3924.
18. Hauke V, et al. (1992) Novel triterpene derived hydrocarbons of the arborane fernane series in sediments: Part II. *Geochim Cosmochim Acta* 56(9):3595–3602.
19. Hauke V, et al. (1995) Isoarborinol through geological times: Evidence for its presence in the Permian and Triassic. *Org Geochem* 23(1):91–93.
20. Jaffe R, Hausmann KB (1995) Origin and early diagenesis of arborinone isoarborinol in sediments of a highly productive fresh water lake. *Org Geochem* 22(1):231–235.
21. Ourisson G, Albrecht P, Rohmer M (1982) Predictive microbial biochemistry—From molecular fossils to procaryotic membranes. *Trends Biochem Sci* 7(7):236–239.
22. Siedenburg G, Jendrossek D (2011) Squalene-hopene cyclases. *Appl Environ Microbiol* 77(12):3905–3915.
23. Fischer WW, Pearson A (2007) Hypotheses for the origin and early evolution of triterpenoid cyclases. *Geobiology* 5(1):19–34.
24. Hoshino T, Sato T (2002) Squalene-hopene cyclase: Catalytic mechanism and substrate recognition. *Chem Commun (Camb)* (4):291–301.
25. Summons RE, Bradley AS, Jahnke LL, Waldbauer JR (2006) Steroids, triterpenoids and molecular oxygen. *Philos Trans R Soc Lond B Biol Sci* 361(1470):951–968.
26. Alain K, Intertaglia L, Catala P, Lebaron P (2008) *Eudoraea adriatica* gen. nov., sp. nov., a novel marine bacterium of the family *Flavobacteriaceae*. *Int J Syst Evol Microbiol* 58(Pt 10):2275–2281.
27. Villanueva L, Rijpstra WI, Schouten S, Damsté JS (2014) Genetic biomarkers of the sterol-biosynthetic pathway in microalgae. *Environ Microbiol Rep* 6(1):35–44.
28. Budzikiewicz H, Wilson JM, Djerassi C (1963) Mass spectrometry in structural and stereochemical problems 32. Pentacyclic triterpenes. *J Am Chem Soc* 85(22):3688–3699.
29. Oyarzun ML, Garbarino JA, Gambaro V, Guilhem J, Pascard C (1987) Two triterpenoids from *Boehmeria excelsa*. *Phytochemistry* 26(1):221–223.
30. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
31. Chen IM, et al. (2016) Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system. *BMC Genomics* 17:307.
32. Banta AB, Wei JH, Welander PV (2015) A distinct pathway for tetrahymanol synthesis in bacteria. *Proc Natl Acad Sci USA* 112(44):13478–13483.
33. Peralta-Yahya PP, et al. (2011) Identification and microbial production of a terpene-based advanced biofuel. *Nat Commun* 2:483.
34. Thoma R, et al. (2004) Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* 432(7013):118–122.
35. Abe I (2014) The oxidosqualen cyclases: One substrate, diverse products. *Natural Products: Discourse, Diversity and Design*, eds Osbourn A, Goss RJ, Carter GT (Wiley Blackwell, Oxford), pp 297–315.
36. Wu TK, Chang CH (2004) Enzymatic formation of multiple triterpenes by mutation of tyrosine 510 of the oxidosqualene-lanosterol cyclase from *Saccharomyces cerevisiae*. *ChemBioChem* 5(12):1712–1715.
37. Lodeiro S, Wilson WK, Shan H, Matsuda SP (2006) A putative precursor of isomalabaricane triterpenoids from lanosterol synthase mutants. *Org Lett* 8(3):439–442.
38. Wu TK, Liu YT, Chang CH, Yu MT, Wang HJ (2006) Site-saturated mutagenesis of histidine 234 of *Saccharomyces cerevisiae* oxidosqualene-lanosterol cyclase demonstrates dual functions in cyclization and rearrangement reactions. *J Am Chem Soc* 128(19):6414–6419.
39. Newman DK, Neubauer C, Ricci JN, Wu CH, Pearson A (2016) Cellular and molecular biological approaches to interpreting ancient biomarkers. *Annu Rev Earth Planet Sci* 44(44):493–522.
40. Pearson A, Budin M, Brocks JJ (2003) Phylogenetic and biochemical evidence for sterol synthesis in the bacterium *Gemmata obscuriglobus*. *Proc Natl Acad Sci USA* 100(26): 15352–15357.
41. Kimura M, Kushiro T, Shibuya M, Ebizuka Y, Abe I (2010) Protostadienol synthase from *Aspergillus fumigatus*: Functional conversion into lanosterol synthase. *Biochem Biophys Res Commun* 391(1):899–902.
42. Lodeiro S, Segura MJ, Stahl M, Schulz-Gasch T, Matsuda SP (2004) Oxidosqualene cyclase second-sphere residues profoundly influence the product profile. *Chem BioChem* 5(11):1581–1585.
43. Meyer MM, Xu R, Matsuda SP (2002) Directed evolution to generate cycloartenol synthase mutants that produce lanosterol. *Org Lett* 4(8):1395–1398.
44. Salmon M, et al. (2016) A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases. *Proc Natl Acad Sci USA* 113(30): E4407–E4414.
45. Li MZ, Elledge SJ (2012) SLIC: A method for sequence- and ligation-independent cloning. *Methods Mol Biol* 852:51–59.
46. Khan SR, Gaines J, Roop RM, 2nd, Farrand SK (2008) Broad-host-range expression vectors with tightly regulated promoters and their use to examine the influence of TraR and TraM expression on Ti plasmid quorum sensing. *Appl Environ Microbiol* 74(16):5053–5062.
47. Amann E, Ochs B, Abel KJ (1988) Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli*. *Gene* 69(2):301–315.
48. Bligh EG, Dyer WJ (1959) A rapid method of total lipid extraction and purification. *Can J Biochem Physiol* 37(8):911–917.
49. Welander PV, et al. (2012) Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology* 10(2):163–177.
50. Summons RE, et al. (2013) Lipid biomarkers in ooids from different locations and ages: Evidence for a common bacterial flora. *Geobiology* 11(5):420–436.
51. Son KC, Severson RF, Arrendale RF, Kays SJ (1990) Isolation and characterization of pentacyclic triterpene ovipositional stimulant for the sweet potato weevil from *Ipomoea batatas* (L.) Lam. *J Agric Food Chem* 38(1):134–137.
52. Nes WD, Wong RY, Griffin JF, Duax WL (1991) On the structure, biosynthesis, function and phylogeny of isoarborinol and motiol. *Lipids* 26(8):649–655.
53. Gaylor JL, Delwiche CV, Swindell AC (1966) Enzymatic isomerization (Delta(8)-Delta(7)) of intermediates of sterol biosynthesis. *Steroids* 8(3):353–363.
54. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
55. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
56. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.
57. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44(Web server issue):gkw290.