Lipid biomarkers: molecular tools for illuminating the history of microbial life

Roger E. Summons[1] ✉, *Paula V. Welander*[2] *and David A. Gold*[3]

Abstract | Fossilized lipids preserved in sedimentary rocks offer singular insights into the Earth's palaeobiology. These 'biomarkers' encode information pertaining to the oxygenation of the atmosphere and oceans, transitions in ocean plankton, the greening of continents, mass extinctions and climate change. Historically, biomarker interpretations relied on inventories of lipids present in extant microorganisms and counterparts in natural environments. However, progress has been impeded because only a small fraction of the Earth's microorganisms can be cultured, many environmentally significant microorganisms from the past no longer exist and there are gaping holes in knowledge concerning lipid biosynthesis. The revolution in genomics and bioinformatics has provided new tools to expand our understanding of lipid biomarkers, their biosynthetic pathways and distributions in nature. In this Review, we explore how preserved organic molecules provide a unique perspective on the history of the Earth's microbial life. We discuss how advances in molecular biology have helped elucidate biomarker origins and afforded more robust interpretations of fossil lipids and how the rock record provides vital calibration points for molecular clocks. Such studies are open to further exploitation with the expansion of sequenced microbial genomes in accessible databases.

Warrawoona Group
A geological unit in Western Australia that includes the remains of microorganisms as old as 3.46 billion years.

Pilbara Craton
An ancient fragment of continental crust in Western Australia that includes rocks as old as 3.6 billion years.

[1]*Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.*

[2]*Department of Earth System Science, Stanford University, Stanford, CA, USA.*

[3]*Department of Earth & Planetary Sciences, University of California Davis, Davis, CA, USA.*

✉*e-mail: rsummons@mit.edu*

https://doi.org/10.1038/s41579-021-00636-2

Life on the Earth was entirely microbial for almost 90% of its history. Microorganisms shaped the world we know and the biogeochemical processes they mediate maintain our planet in a state that is habitable for complex, multicellular life. Photosynthesis by cyanobacteria and algae produces most of the oxygen we breathe, nitrogen fixation by free-living and symbiotic microorganisms provides an essential nutrient, organic matter remineralization recycles what is no longer living, rock weathering delivers essential nutrients to aquatic environments and a myriad of other reactions cleanse those same waters and the atmosphere of toxic substances[1]. Still, the rock record holds surprisingly little direct evidence of this microbial world, especially during its earliest stages.

Considering the vastness of the continental land masses and sea-floor sediments that record life's early history, just a sliver can be dated to the Archaean aeon (4,000–2,500 million years ago (Ma)) and none, other than a few grains of the mineral zircon[2], remain from the preceding Hadean aeon (~4,560–4,000 Ma). Although there are no direct records of when, where or how life originated, reconstructions that integrate genomic and fossil data place this process to within the Hadean aeon[3] and, possibly, as long as 1 billion years prior to the oldest signatures of life identified in the rock record. Widely accepted evidence for a biologically active Archaean Earth occurs in the ~3,470–3,450 Ma Warrawoona Group[4] of the Pilbara Craton, Western Australia, and includes assemblages of carbonaceous microstructures[5–7], stromatolites[8–10] and chemical signals[11] that confirm the existence of functioning ecosystems (FIG. 1). Unlike other ancient terranes, these comparatively well-preserved rocks were deposited, close to sea level, on the oldest known block of emergent continental crust[12]. Carbon isotopic data for limestones and organic carbon phases, combined with modelling approaches, suggest the existence of a stable biogeochemical carbon cycle[11], a temperate climate and oceans of near neutral pH during this same time period. Isotopic data infer microbial cycling of sulfur[13,14] and nitrogen[15] as well. However, no fossils, physical or chemical, dating from these oldest of rocks preserve sufficient detail to reconstruct the specific identity of the earliest microorganisms, other than isotopic inferences for the existence of the physiologies of photosynthetic carbon fixation, respiration, nitrogen fixation and sulfate reduction.

Recognition of a radically different type of fossil, distinct from traditional body fossils, came with the inception of the new field of molecular evolution largely flowing from the collaboration between Emile

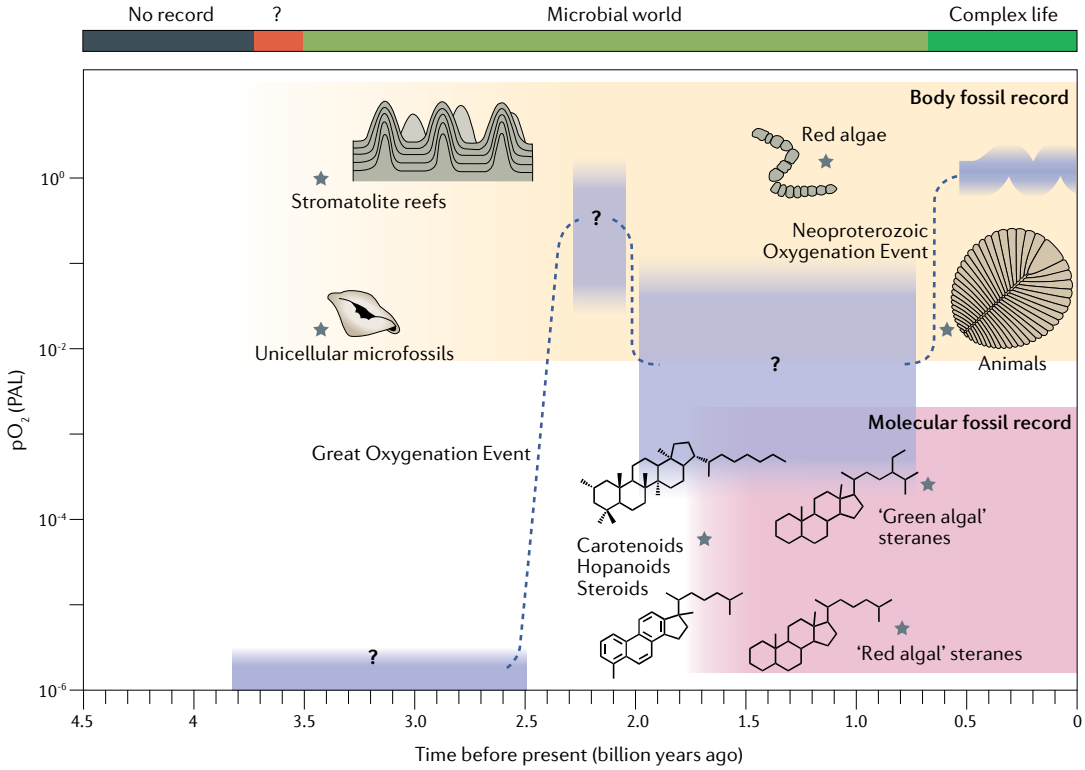


Fig. 1 | **Geological timescale of the Earth's fossil record in the context of planetary oxygenation.** The bar at the top summarizes how sedimentary rocks record biological evolution: black represents time interval where there is no rock record, red where there are disputed records, light green where there are records of exclusively microbial life and dark green where there are records of complex animal and plant life. The areas shaded mauve and the dashed lines depict the progressive rise in atmospheric oxygen as presently understood from diverse geochemical proxies[140,141]. The area shaded pink depicts the time when there is a robust record of molecular fossils. Stars identify the oldest known occurrences of the named chemical and physical fossils. PAL, present atmospheric level; pO_2, partial pressure of atmospheric oxygen.

Zuckerkandl and Linus Pauling in the 1960s (REF.[16]). In the initial concept, proteins by way of their amino acid sequences served as "documents of evolutionary history"[17]. As protein sequences changed at perceptible rates, the concept of a molecular clock was born and soon expanded to include the three informational macromolecules of protein, RNA and DNA[18]. However, all three biopolymers are easily hydrolysed, imposing severe limitations on their preservation and restricting study of ancient counterparts[19] apart from the study of reconstructed analogues[20,21]. Although more limited in information content, the hydrocarbon cores of microbial membrane lipids also serve as molecular fossils. Extremely recalcitrant once entombed in sedimentary rocks, these can be preserved on billion-year timescales under ideal circumstances[22]. Such lipids, termed 'biomarkers' as a contraction of 'biological marker compounds'[23,24], are biosignatures for particular classes of organisms according to their taxonomic specificity. It is for this reason that studies of fossilized biomarkers have provided insight into a range of phenomena including the transitions in ocean phytoplankton through time[25], exemplified by the apparent switch from oceans dominated by bacteria to those where algae assumed major significance[26]. Biomarkers have also proved exceptionally useful in revealing the redox structure of seas during mass extinction events and other periods of biotic turnover[27,28], extreme ocean temperature regimes in the past[29,30] and the identity of otherwise enigmatic fossils[31]. Lipid biomarkers, because of their preservation in ancient sediments, can also serve as calibration points for molecular clocks. This is especially crucial for ancient microbial life where the record of recognizable body fossils is sparse and where the majority of forms are described as acritarchs, meaning organic microfossils with unknown biological affinities[32].

Biomarker research was founded on observations from the field of natural products chemistry[33] and was initially conducted by chemists endeavouring to understand the molecular character of petroleum and organic substances preserved in rocks and in fossils themselves[34,35]. As a result, molecules with previously uncharacterized structures were often identified in the geological record well before their biological precursors were discovered in contemporary organisms. The prime example, hopanoids, were recognized in ancient sediments and petroleum well before their biological counterparts were identified in extant bacteria[36,37]. In another example, a C_{20} 'highly branched isoprenoid' hydrocarbon with a propeller-like branching pattern was identified in a petroleum seep near the Great Salt Lake, USA, and its structure established by spectroscopy and chemical synthesis[38]. This soon led to discovery of analogous C_{25} and C_{30} hydrocarbons[39–41] along with an

Stromatolites
Layered sedimentary structures that form when a microbial community traps and binds sediment grains.

Hopanoids
A class of molecules comprising six C_5 isoprene units folded into a pentacyclic ring system.

appreciation for their prevalence across different environments. However, it was another decade before the source was revealed in cultures of diatoms[42] and the geological, environmental and evolutionary significance of these unusual hydrocarbons, now known as haslenes and rhizenes, was appreciated[43,44]. These case studies revealed a pressing need for additional systematic approaches to characterize the lipids in extant organisms, to better inform interpretations of biomarker signals in the sedimentary rock record[45–48]. However, every success story raises new questions. For example, what physiological function do these highly branched C_{25} and C_{30} isoprenoid hydrocarbons perform in diatoms? How are they biosynthesized and when did the pathway originate? Is it limited to diatoms and, if so, which species? Can they be used to track specific plankton communities or the conditions under which they thrive? In this Review, we explore recent work that has been done to link lipid biomarkers to their biological sources and demonstrate how genomics has opened the way to address many remaining questions, including those concerning the nature and history of early microbial life.

Lipids as biomarkers

Although lipids are ubiquitous in environmental and geological materials, only a fraction of them function as biomarkers. The term 'lipid' covers a large array of biomolecules that are operationally defined by their insolubility in water and solubility in non-polar solvents, and which encompass a diverse suite of compound classes from simple hydrocarbons to fatty acids, alcohols and pigments. The fatty acids and alcohols can be ester or ether linked to glycerol, which can be further modified with polar head groups to form the amphiphilic lipid bilayer that surrounds most cells. Hydrocarbon chain lengths, how chains are linked to glycerol (ester versus ether), degrees and sites of unsaturation and classes of polar head group (phosphate, amino functions, sugars and so on) all have relationships to phylogeny and microbial community structure[49–51]. However, much of this information is lost to hydrolysis or oxidation after an organism dies[52,53]. Rather, the non-hydrolysable hydrocarbon cores of complex membrane lipids, as well as the structurally diverse polyisoprenoids, have the most secure connections to physiology, phylogeny and environmental conditions[48,54]. Commonly encountered lipid classes (FIG. 2) that have biomarker potential include acyclic isoprenoids, sterols, bacteriohopanepolyols, glycerol dialkyl glycerol tetraethers (GDGTs) and carotenoids[55–57]. So long as ancient sediments have not been excessively heated, the diagnostic structural features of lipids are still recognizable, despite the changes wrought by diagenesis and catagenesis[58].

Lipid biomarkers and genomics

The interpretation of lipid signatures preserved in ancient settings requires an understanding of the occurrence of their precursors in extant organisms, what their physiological roles might be and how their production, function and preservation might be influenced by environmental factors. Rigorous lipid surveys over the past 50 years have provided crucial information regarding the organisms that synthesize these lipids as well as the types of environments in which they are abundant[47,59,60]. These traditional lipid analysis studies have been foundational to the biomarker field and formed the solid ground from which all biomarker work has expanded — including recent molecular work combining comparative genomic analyses with genetic and biochemical studies. These genomic and biochemical approaches have revealed previously unknown proteins required for the synthesis of various biomarker lipids[61–63] that have been used to search large genomic and metagenomic databases for the potential to produce these lipids in untested organisms and directly in specific environments[64–67]. The first study to demonstrate the utility of this approach used the squalene–hopene cyclase sequence to identify novel hopanoid cyclases in environmental metagenomic data sets[64], demonstrating that our knowledge of biomarker lipids in modern systems can be expanded through the application of genomic approaches[60].

There are several benefits to the use of comparative genomics in addition to traditional lipid analyses in biomarker studies (BOX 1). Firstly, genetic data are accumulating at a rate far faster than lipid data. The number of nucleotide bases in GenBank (a DNA sequence repository run by the US National Institutes of Health) has doubled approximately every 18 months since 1982, with more than 1 billion sequences accessible for public research[68]. Scientists can leverage this massive data set to infer which organisms carry the genes necessary to synthesize lipids of interest. Metagenomic data sets — genetic data collected directly from environmental samples — have proven particularly valuable, as the genes and organisms responsible for biomarker synthesis can be inferred even when the species in question prove difficult to culture in the laboratory. A second benefit of genetic data is that a suite of tools and techniques exist to study how genes evolve over time. Although the presence or absence of lipids in living organisms provides clues into biomarker sources, such data are insufficient to determine when organisms evolved the ability to synthesize biomarkers, or whether such biomarkers were lost in some lineages in the past. By contrast, bioinformatic tools can be used to reconstruct histories of gene gain, loss and duplication, allowing scientists to infer the lipid repertoires of extinct lineages. Finally, understanding the geological significance of the new discoveries flowing from the revolution in genomics requires contextualizing biomarker genes in a phylogenetic (evolutionary) framework. Genes can have evolutionary histories distinct from the organisms that house them, meaning that the distribution of genes among living species can create a false impression of a biomarker's origin. Although 'gene trees' may therefore diverge from species trees that are based on established phylogenetic markers (16S rRNA or concatenated protein sequences) (BOX 2), they can provide a sense of the distribution of genes across taxa and snapshots of the evolution of biomarker biosynthesis. Molecular clock approaches, which utilize fossil data to calibrate phylogenetic trees, can then be used to constrain the timing of important evolutionary events (BOX 2). In the following section, we explore a series of case studies where the methodology

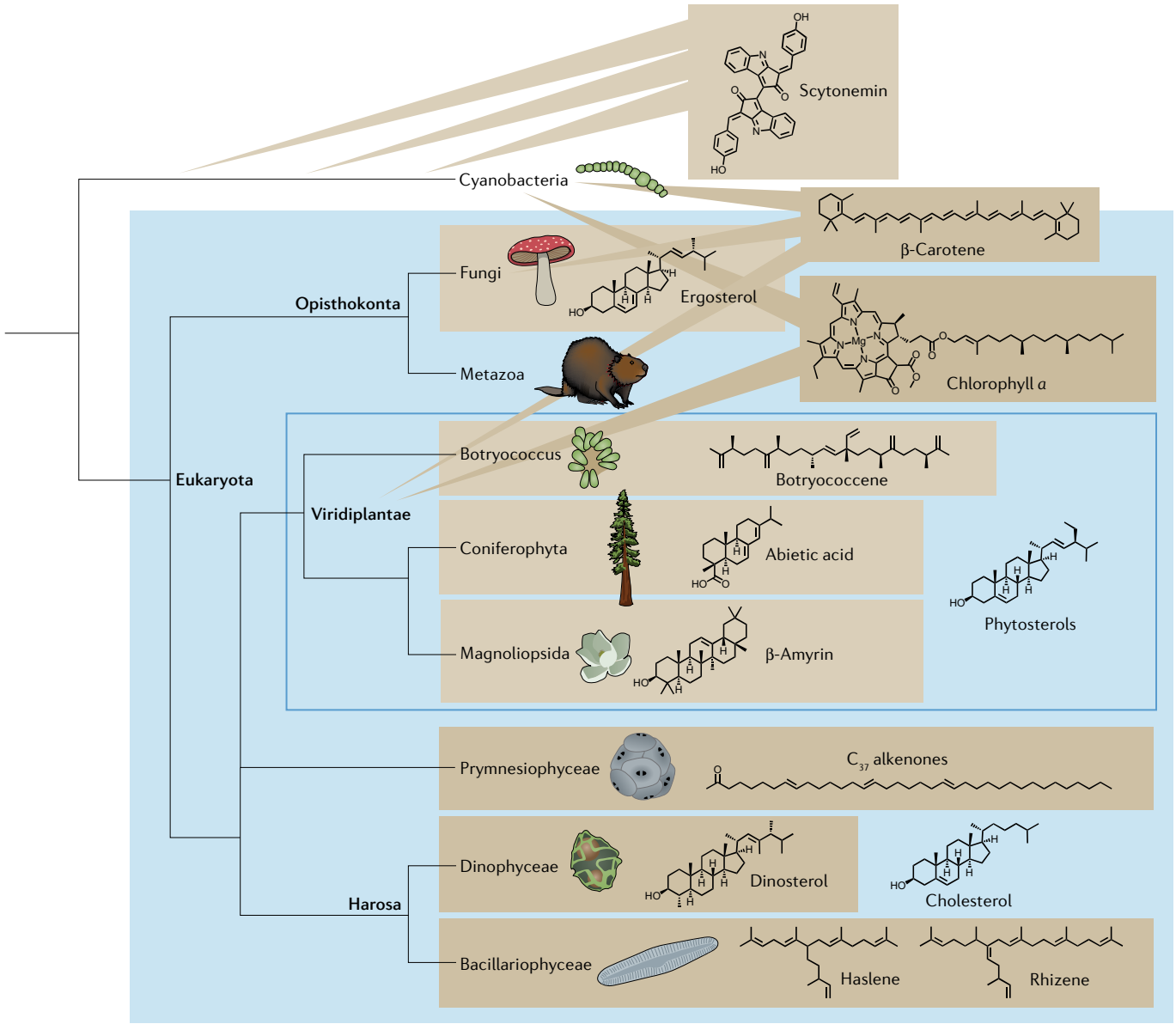


Fig. 2 | **Commonly referenced lipid biomarkers and their known sources.** Illustration of how lipid biomarkers can map to organismic phylogenies. Boxes encompassing different branches indicate broad distributions, whereas multiple arrows emanating from different branches indicate likely evolutionary convergence between groups, and arrows emanating from the same branch (for example, scytonemin) signify a biomarker that evolved multiple times within one group. Some classes of biomarkers (for example, alkenones, botryococcenes, haslenes and rhizenes) consist of groups of molecules that differ only in carbon number, degree of unsaturation and so on, but otherwise share close structural similarities. These similarities imply biosynthetic pathways specific to each class.

described above has helped elucidate the origins of geochemical biomarkers, as well as issues ripe for future investigations.

Genomics-enabled advances in biomarkers

Fossilized carotenoid pigments as biomarkers. The saturated bicyclic hydrocarbon carotane was one of the earliest examples of a sedimentary biomarker characterized using combined gas chromatography–mass spectrometry[69]. The presumed precursor, β-carotene, occurs widely as an accessory pigment in plants, algae and cyanobacteria, rendering carotane a biomarker for chlorophototrophs but a relatively non-specific one. More limited in occurrence, and representing very specific ecological niches, are the aromatic carotenoids isorenieratene and chlorobactene of the green sulfur bacteria[70], okenone derived from the purple sulfur bacteria[71] and their fossilized counterparts[72]. Both classes of anaerobic bacteria require sulfide as an electron donor for photosynthesis but differ in the quality of light utilized and tolerance of oxygen[73]. Different dispositions of methyl groups on the rings of these aromatic carotenoids (1-alkyl-2,3,6-trimethyl (ϕ-ring) in the case of the carotenoids isorenieratene and chlorobactene, and 1-alkyl-2,3,4-trimethyl (χ-ring) for the carotenoid okenone) means that the two types of microbial phototroph can

Accessory pigment
A coloured molecule that absorbs light and works in concert with the primary pigment, typically chlorophyll *a*.

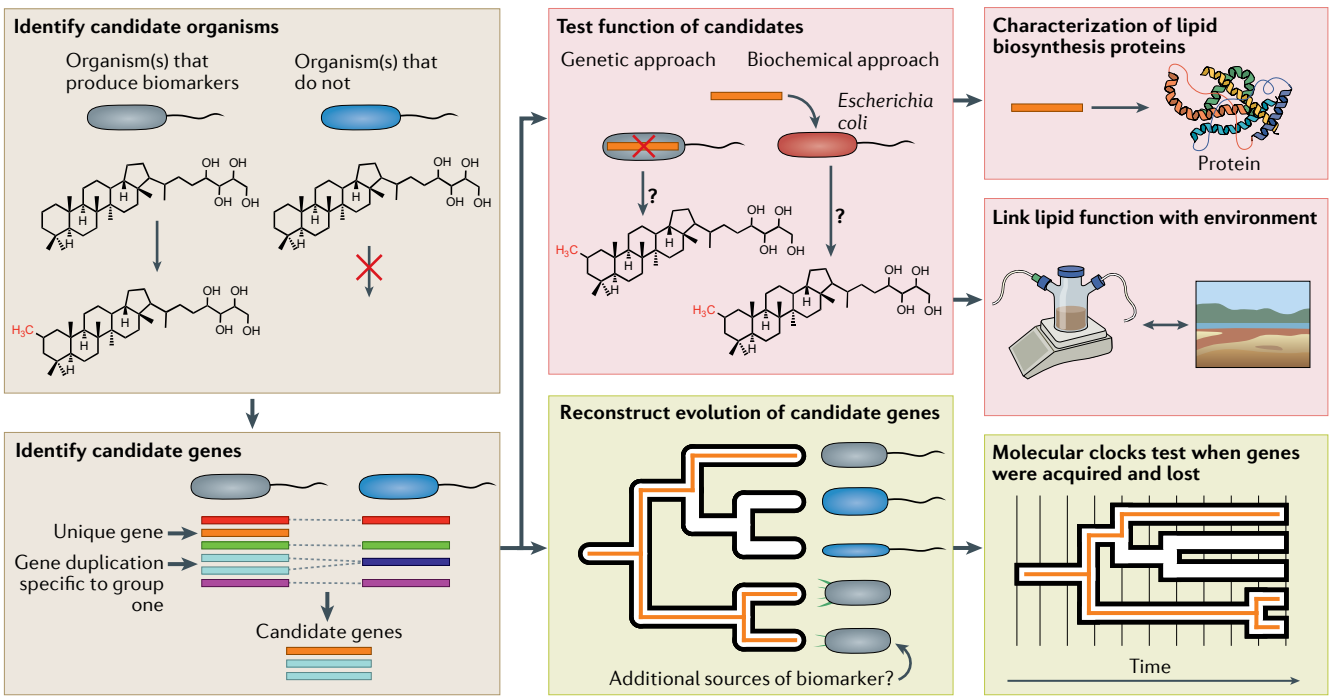
be readily distinguished in gas chromatography–mass spectrometry analyses[22,74,75].

Genomics and mutagenesis studies of carotenoid biosynthesis have revealed additional complexities while also revealing a clear path to decoding fossil carotenoid origins. Aromatic carotenoids form through dehydrogenation of carotene rings accompanied by rearrangement of the methyl substituents. In *Chlorobium tepidum*, this process is catalysed by a γ-carotene-desaturase/methyltransferase denoted CrtU to form chlorobactene, a monoaromatic[76]. In brown strains of the green sulfur bacteria such as *Chlorobaculum limnaeum*, the final product is the diaromatic isorenieratene[77,78]. In both processes, the result is aromatic carotenoids with the 1-alkyl-2,3,6-trimethyl ϕ-ring. By contrast, the CrtU protein in Chromatiaceae affords carotenoids with the 1-alkyl-2,3,4-trimethyl χ-ring[79]. Further, genomics analyses revealed a homologous gene in some actinobacteria consistent with their production of isorenieratene[80,81] and in cyanobacteria. However, at the time, the function of CrtU in cyanobacteria was not known as cyanobacteria were not known to produce aromatic carotenoids[76]. Subsequently, culture studies and pigment analyses revealed that a protein, now denoted CruE, converts β-carotene to renierapurpurin, an aromatic carotene with dual χ-rings, and that this is

Box 1 | **Identifying lipid biomarker biosynthesis genes and proteins in extant organisms**

The use of comparative genomics to identify candidate biomarker biosynthesis genes in extant organisms has proven to be one of the most effective approaches in narrowing down the biological sources of lipid biomarkers[64,142]. This approach entails identifying one or more organisms that produce a biomarker of interest and a set of organisms that do not. A comparative analysis of the genomes between these two sets of organisms (lipid producers versus non-producers) will return a set of candidate genes that are targeted for further study. These gene sets can range in size from manageable (for example, 20 genes) to unmanageable (for example, 2,000 genes). If gene sets are quite large, then a second set of comparisons can be executed to take advantage of any knowledge of the potential chemistry required to synthesize lipids of interest. For example, early experiments with labelled methionine suggested that the synthesis of 2-methylhopanoids involves *S*-adenosylmethionine activity, a common cofactor in many biochemical reactions[143]. Thus, the candidate gene list for potential C-2 hopanoid methylation could be further sifted to focus on proteins that were predicted to have an *S*-adenosylmethionine-binding motif[142]. Tools such as basic local alignment search tool (BLAST)[144] or HMMER[145] combined with the Pfam database[146] can help determine whether predicted binding sites are present in a candidate gene set. One can then experimentally determine whether a candidate gene encodes for a lipid biosynthesis protein through either genetic or biochemical experimental approaches. The genetic approach requires the disruption of the candidate gene, either through a replacement of the gene locus with an antibiotic marker ('gene disruption') or a complete deletion of the gene locus ('markerless gene deletion'). The biochemical approach requires the expression of the gene of interest in model laboratory systems, such as the bacterium *Escherichia coli* or the yeast *Saccharomyces cerevisiae*, and then carrying out biochemical assays with cell-free extracts or purified protein to demonstrate that the candidate protein catalyses the lipid biosynthesis reaction of interest.

The identification of specific lipid biosynthesis genes opens other avenues for further research that can improve understanding of biomarker lipids. The identification of biosynthesis genes allows the monitoring of gene expression through in vivo transcriptional analyses both in cultured organisms and directly in different, contemporary ecosystems. These studies, coupled with biochemical characterization of the enzymes encoded by these lipid biosynthesis genes, would provide insight into how lipid production is influenced by specific environmental factors. In addition to the direct characterization of lipid biosynthesis enzymes, it is also possible to generate gene deletion mutants that are no longer capable of producing lipids of interest. Physiological studies of these mutants provide insight into the conditions under which synthesis of lipid biomarkers is essential and would reveal what functional role these lipids have in different cells[147–150]. The figure shows an example of how biosynthesis proteins, in this case the one associated with the formation of 2-methylbacteriohopanepolyols, are identified in biomarker-producing microorganisms through the use of comparative genomics, gene deletion analyses and heterologous expression.

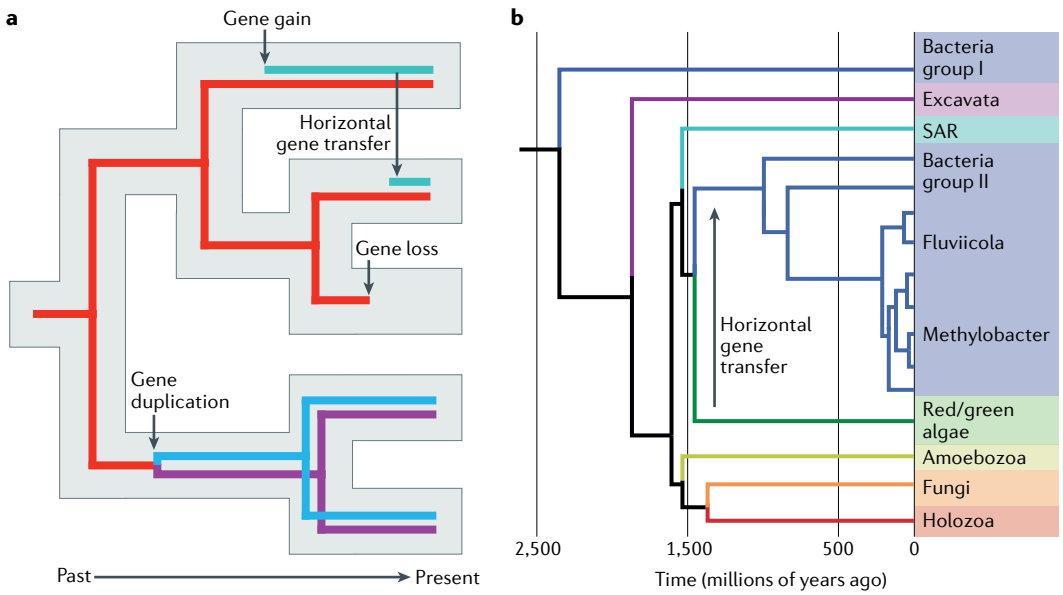


Box 2 | **Gene trees versus taxon trees**

Three processes cause gene trees to diverge from a species tree: gene gain — including de novo origination as well as horizontal gene transfer — gene loss and gene duplication (see the figure, part **a**). Some genes involved in lipid biosynthesis are rarely gained, lost or duplicated, but in others these processes are commonplace. For genes with complex evolutionary histories, typical tree-building software is unlikely to produce an accurate topology. Gene duplications are particularly troublesome, as evolutionary constraint is often relaxed on one of the two paralogs (a process known as neofunctionalization), leading to increased rates of sequence mutation, poor node support and long branch artefacts[151]. Programs such as NOTUNG[152] and TreeFix[153] can help by using the known relationships between species to guide the production of a more accurate gene tree.

Once a gene tree has been constructed, molecular clock approaches can further elucidate the timing of important events (see the figure, part **b**). Molecular clocks are a suite of computational analyses that use variation in gene sequences to estimate divergence times in phylogenetic trees. In other words, they convert the relative time inherent in a phylogeny (for example, *Methylobacter* spp. did not obtain sterol biosynthesis genes until after green algae evolved) into absolute time (for example, *Methylobacter* spp. obtained sterol biosynthesis genes ~400 million years ago). The accuracy of a molecular clock is driven by the amount and quality of genetic data, as well as the proper use of fossil data to 'calibrate' the tree[154]. Both of these aspects can be problematic for microbial gene trees; the number of informative sites in a gene is limited by its size, and the microbial fossil record is limited and highly debated[155,156]. Because of these limitations, claims of exact dates or ages produced by molecular clocks should be treated with caution. Nevertheless, a molecular clock approach can be powerful for adjudicating between competing hypotheses of biomarker evolution.

The figure depicts relationships between gene trees and taxon trees. Part **a** is a hypothetical construction showing how an ancestral gene might become distributed in descendants by gain, loss or duplication. Part **b** is a simplified tree of the specific case of the gene coding for the squalene monooxygenase (SQMO) protein involved in sterol biosynthesis in eukaryotes and its distribution in two distinctly different clades of bacteria. The SAR supergroup is a clade of protists comprising the stramenopiles, alveolates and Rhizaria. Figure part **b** adapted from REF.[101], Springer Nature Limited.



further transformed to synechoxanthin, a highly polar χ,χ-caroten-18,18′-dioic acid[82] in a series of oxidative processes thought to be catalysed by the enzyme CruH[83]. Thus, microorganisms that harbour both *cruE*/*crtU* and *cruH* are candidates for production of synechoxanthin and basic local alignment search tool (BLAST) searches reveal their simultaneous presence in diverse cyanobacteria (FIG. 3). This series of observations, therefore, illuminates the reasons behind some previously enigmatic geochemical observations[84,85].

The carotenoids renierapurpurin and renieratene, as their names imply, were first isolated and characterized in extracts of the sea sponge *Reniera japonica*[86,87]. However, those sponges whose genomes have been sequenced are not known to contain the genes necessary for biosynthesis of aromatic carotenoids. By contrast, cyanobacteria are common amongst the ~30 clades of bacteria identified in sponge microbiomes[88] and are likely a major source of the carotenoids found in sponges. Accordingly, cyanobacteria are also candidate sources that account for the presence of renierapurpurane and renieratane in the oldest sedimentary rocks that have yielded fossil hydrocarbons[22] as well as in numerous more recent marine and non-marine sediments where these carotenoids have been identified[89,90]. Lastly, lipids with aromatic carboxylic acid functional groups are prone to decarboxylation during the early stages of sedimentary diagenesis. This predicts that synechoxanthin and the monocarboxylic acid biosynthetic precursors should be transformed into C_{38} and C_{39} diagenetic product hydrocarbons upon reduction and burial[90,91]. Thus, cyanobacteria are now a valid source for carotenoids on the biosynthetic pathway to synechoxanthin including compounds previously attributed to anoxygenic phototrophs[85,92,93].

Sedimentary diagenesis
The processes by which sedimentary rocks and their components became modified over time during burial.

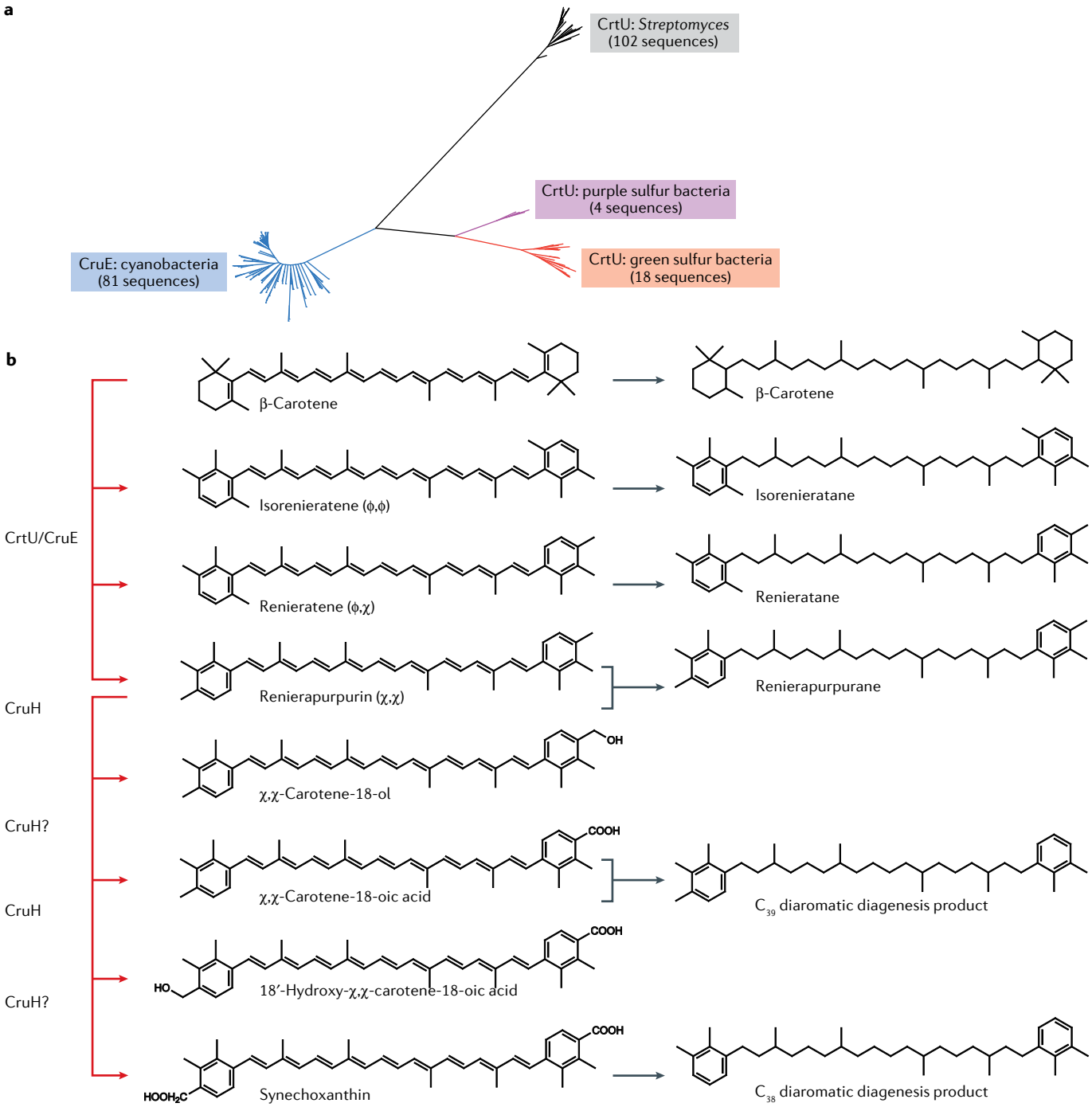


Fig. 3 | **Phylogeny of the proteins essential for biosynthesis of aromatic carotenoids. a** | Phylogenetic tree showing distribution of CrtU and CruE proteins across sequenced bacteria, demonstrating the clustering of these proteins into four distinct clades of green sulfur bacteria, purple sulfur bacteria, *Streptomyces* and cyanobacteria. Further basic local alignment search tool (BLAST) searches reveal that only cyanobacteria appear able to produce synechoxanthin. **b** | Biosynthetic pathways of some carotenoids discussed in the main text and structures of their fossilized hydrocarbon counterparts.

Sterol origins, early eukaryotes and planetary oxygenation. Steranes are an example of fossil biomarkers whose origins have been greatly informed by genomic data. Steranes are derived from sterols, a class of lipids produced by all major groups of eukaryotes. The oldest confirmed sterane occurrence is found in the ~1.65 billion-year-old Barney Creek Formation of Australia[22]. Unfortunately, conflicting lines of data make it difficult to interpret these biomarkers. Barney Creek steranes appear to be primitive compared with the structurally diverse nature of contemporary eukaryotic sterols. They lack alkylation at C_{24} and include triaromatic 4-methyl steranes in relatively high abundance, suggesting a truncated biosynthetic pathway[85]. More conventional sterane distributions predicted to represent eukaryotic sterols are not found until substantially later in the Neoproterozoic era, ~800 Ma,

with no confirmed occurrences in the interim[94]. But this too is problematic, as there are well-accepted eukaryote fossils greater than 1 billion years old[95,96]. Further complicating matters, a few bacteria have long been known to make simple sterols, and more recent molecular data demonstrated that sterol biosynthesis genes occur in diverse bacteria[97,98]. Determining when these various groups evolved the ability to synthesize sterols is therefore integral for interpreting the sterane fossil record and eukaryote origins.

To address these perplexing observations, one study used a molecular clock approach to test when different species evolved the genes necessary to build basic sterols. Sterol biosynthesis begins with two enzymes — squalene monooxygenase (SQMO) and oxidosqualene cyclase (OSC). These two enzymes convert the acyclic polyisoprenoid lipid squalene into a tetracyclic protosterol, either lanosterol or cycloartenol. Querying genetic databases, the authors observed that both SQMO and OSC show almost identical evolutionary patterns, with bacterial sequences clustering into two groups. Because *sqmo* and *osc* are both members of larger gene families, out-group genes could be used to polarize the evolutionary tree and determine when these bacterial clusters originated (BOX 2). One group was nested within the eukaryotes, suggesting horizontal gene transfer from eukaryotes to bacteria, whereas the other group fell outside the eukaryotes, suggesting horizontal gene transfer between bacteria and the ancestor of living eukaryotes. Although there is substantial uncertainty in the timing of this first horizontal gene transfer event, the most likely timing for this event is around ~2.3 billion years ago, around the same time that the geological record shows convincing evidence for the irreversible atmospheric oxygenation — The Great Oxygenation Event[99,100] (FIG. 1). This suggests that bacteria and proto-eukaryotes were swapping genes necessary for sterol biosynthesis, an oxygen-intensive process, around the same time that oxygen became a permanent gas in the atmosphere and shallow ocean. Importantly, one study[101] also refutes the hypothesis that modern sterols have a Neoproterozoic (1,000–541 million years ago) origin but, rather, states that the observed patterns[94] represent the ecological expansion of eukaryotic plankton.

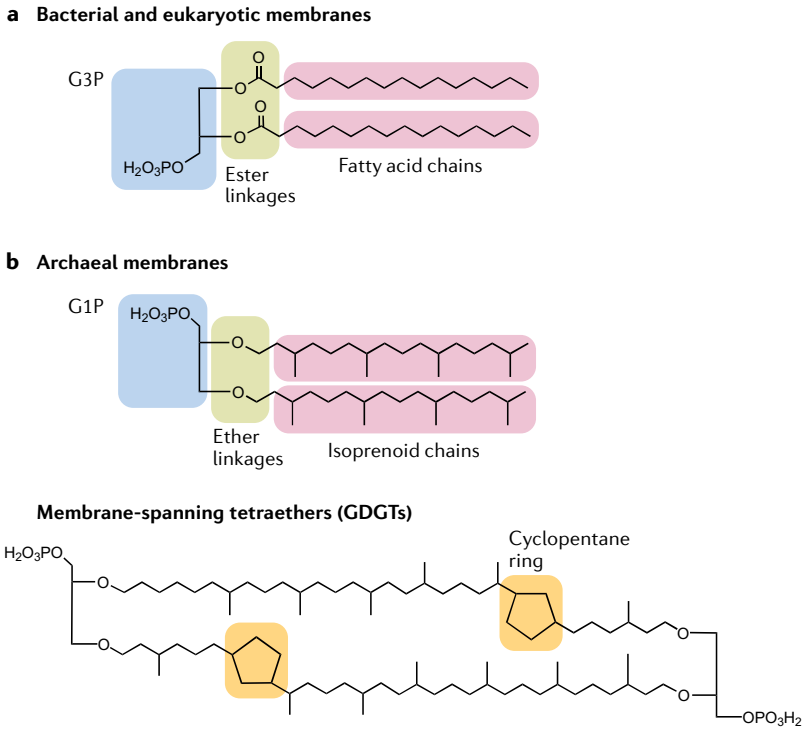


Fig. 4 | **Structural differences between archaeal and bacterial membrane phospholipids. a** | Bacterial and eukaryotic membranes are composed of bilayers of fatty acid lipids ester linked to glycerol-3-phosphate (G3P). **b** | Archaeal membranes are composed of isoprenoidal lipids ether linked to glycerol-1-phosphate (G1P) and bilayers can be fused to form membrane spanning lipids with distinctive cyclopentane rings. Pink, green and blue represent the three characteristics that differentiate archaeal membrane lipids from bacterial and eukaryotic membrane lipids: archaea have isoprenoid lipid chains in contrast to fatty acid chains in bacteria and eukaryotes (pink); archaeal isoprenoid chains are ether linked to the glycerol backbone whereas bacterial and eukaryotic fatty acid chains are ester linked to the glycerol backbone (green); and stereochemistry of the glycerol backbone is distinct between the two classes of lipids (blue). Archaea link their lipids to G1P whereas bacteria and eukaryotes utilize G3P. GDGT, glycerol dialkyl glycerol tetraether.

Biomarkers for archaea and palaeotemperature assessments using archaeal lipids. Although archaea, particularly methanogenic and halophilic archaea, have been studied for almost 100 years[102], it was only ~44 years ago that molecular phylogenetic analyses revealed that these microorganisms represent a domain of life separate from bacteria[103]. More recently, phylogenomics has revealed an evolutionary relationship between eukaryotes and archaea[104]. Initially, archaea were thought to be primarily extremophilic and limited in ecosystem diversity, but environmental genomic studies indicate that these microorganisms inhabit diverse environments and can be metabolically flexible[104]. In addition, the biogeochemical impacts of archaea are proposed to be quite significant both in modern environments and historically. However, the origins and evolution of archaea have been difficult to pinpoint as biomarker evidence for archaea in deep time is sparse[105].

The sparsity of archaeal lipid biomarkers is partly a reflection of their membrane chemistry. To date, no polycyclic triterpenoids such as sterols or hopanoids have been identified in archaeal cultures[106]. The lack of polycyclic terpenoids, which function as rigidity modulators in bacterial and eukaryotic membranes, might indicate that archaea have alternative mechanisms for modulating membrane behaviour[106]. This would not be surprising, as one of the key features that distinguishes archaea from bacteria and eukaryotes is the chemical nature of their cellular membranes. Whereas bacterial and eukaryotic membranes are composed of fatty acid chains ester linked to glycerol-3-phosphate (G3P), archaea membranes are composed of isoprenoid-based hydrocarbon chains ether linked to glycerol-1-phosphate (G1P)[107] (FIG. 4). In addition, many, but not all, archaea generate a tetraether membrane-spanning monolayer of GDGTs that can be further modified by the addition of cyclopentane or cyclohexane rings. These unique archaeal diether and tetraether isoprenoid-based membrane lipids can function as biomarkers indicating the occurrence of archaea in sedimentary records. Diagenetic products

of archaea membranes include regularly branched acyclic isoprenoids up to C_{25} from archaeols and extended archaeols, C_{40} biphytanes from GDGTs[108] and C_{80} isoprenoids from glycerol monoalkyl glycerol tetraethers[109]. Acyclic isoprenoids with 25 carbon atoms have been identified in ancient sedimentary rocks dating as far back as 1.6 billion years[110]. However, these archaeal markers are typically not diagnostic beyond the domain level — they may indicate the occurrence of archaea but do not denote specific archaeal taxa or metabolism.

Whereas the acyclic phytanes and biphytanes may only serve as domain-level biomarkers, the core GDGTs have the potential to be informative not as biomarkers for specific archaeal taxa but, rather, as indicators of past environmental conditions through their use as palaeotemperature proxies[57]. Palaeoclimatology is the reconstruction of past temperature changes through the use of various proxies including the chemical and isotopic make-up of skeletal carbonates and the distribution of microbial lipid biomarkers preserved in ancient sediments[111]. Lipid-based proxies are particularly useful as they may be selectively preserved in some sedimentary environments, because they complement or supplement other indicators, or because they afford reconstructions of past temperature changes deeper in geological time[111]. Lipid-based palaeotemperature proxies have been developed over the years based on the membrane lipids of microalgae[29], bacteria[112] and archaea[30,113]. The TEX_{86} (tetraether index of 86 carbons) palaeotemperature proxy was the first palaeotemperature proxy that utilized a ratio of preserved membrane lipids specific to archaea — the core GDGTs[114]. Core GDGTs are subject to diagenetic degradation under high temperature and/or oxic conditions, and are unlikely to be well preserved in materials that have been exposed to high levels of geothermal heating[115]. However, they do occur in sedimentary rocks dating as far back as the middle Cretaceous (~100 million years ago) and are useful to reconstruct changes in atmospheric and sea surface temperatures (SSTs) over more recent geological timescales[116,117].

The TEX_{86} SST proxy was first established as a ratio of the relative abundance of specific cyclized GDGTs that correlated with annual mean SST in surface sediment data[114]. Subsequent calibration of the TEX_{86} proxy with more extensive sediment data sets has demonstrated limitations for the application of this proxy in certain environments (for example, polar oceans and large lakes)[118–120]. Nonetheless, various environmental studies continued to demonstrate that TEX_{86} and related proxies correlate well to SST in marine systems[57,121,122]. Yet there are some underlying assumptions that are made in the application of TEX_{86} that undermine its robustness. In particular, laboratory studies of pure archaeal cultures demonstrate that factors other than temperature, such as changes in pH, oxygen availability or nutrients, can impact GDGT cyclization[123–125]. It is also proposed that one group of archaea, the Marine Group I (MGI) Thaumarchaeota, are the dominant source of cyclized GDGTs in the surface ocean[111] whereas metagenomic studies have demonstrated that uncultured MGII Euryarchaeota dominate the surface oceans and are, therefore, another potential source[126,127]. This is significant because changes in cyclized GDGTs observed in surface sediments may reflect changes in archaeal communities, rather than a response to temperature fluctuations, and this could distort TEX_{86} temperature assessments.

One obstacle to constraining these competing hypotheses is that the biochemical mechanism of cyclization — and the proteins that carry out this biochemistry in the Thaumarchaeota — are largely unknown. A recent study identified two GDGT ring synthases, termed GrsA and GrsB, in the thermoacidophile *Sulfolobus acidocaldarius*[128]. GrsA and GrsB are radical *S*-adenosylmethionine proteins that are responsible for adding cyclopentane rings at the C-3 and C-7 positions, respectively, in the core GDGT structure (FIG. 5). Using these two proteins as search queries, it was possible to mine metagenomic data from the North Pacific and confirm that the MGI Thaumarchaeota were the only potential source of cyclized GDGTs in the water column — directly constraining one uncertainty of the TEX_{86} proxy. A recent lipid study also concluded that MGII did not contribute significantly to the cyclized GDGT pools of the North Atlantic Ocean and the coastal North Sea[129]. These data suggest that MGII are not producing cyclized GDGTs and have led to the hypothesis that these archaea may harbour distinct membrane lipids. Future work can now focus on characterizing the biochemical details of GDGT ring formation by the Grs proteins[130] and how their expression is affected by various environmental factors, including temperature.

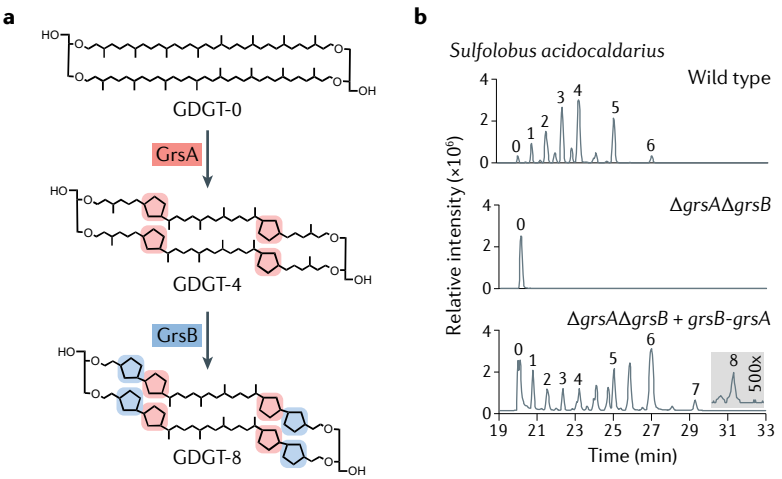


Fig. 5 | **Identification of GDGT ring synthase proteins in *Sulfolobus acidocaldarius*.** **a** | Cyclization reactions catalysed by the two glycerol dialkyl glycerol tetraether (GDGT) ring synthase (Grs) enzymes. **b** | Liquid chromatography–mass spectrometry chromatograms showing distributions of cyclized GDGTs in wild-type lipid extracts (top). Deletion of *grsA* and *grsB* results in the loss of cyclized GDGTs (middle) but production is restored when *grsA* and *grsB* are introduced into the mutant strain on a plasmid (bottom). Numbers above each peak indicate number of rings in the GDGT structure. Figure adapted with permission from REF.[128], PNAS.

Progress, pitfalls and future prospects

The field of biomarker genomics has seen tremendous progress, yet many issues and questions remain. Contamination has been a persistent challenge in biomarker research, especially since the mass spectrometry

Carbonaceous chondrites
Carbon-rich meteorites composed of small mineral grains and representing some of the post-primitive material in the solar system.

methods used to detect lipids have become exquisitely sensitive. Different types of contamination occur in geological versus biological samples. Biomarkers can leach into rocks from surrounding sediments or from petroleum seepage. Drilling cores into rocks can introduce hydrocarbons into freshly exposed samples, whereas aerosols contaminate museum specimens and other curated materials over time[131,132]. Techniques to identify authigenic biomarkers and limit anthropogenic contamination include comparing the surfaces with interiors of geological samples, checking out any anomalous signals or having the same sample tested in multiple independent laboratories. Biological samples can be compromised by biomarkers from elsewhere in the environment or from handling. Genome sequencing can also hint at contamination; if an organism lacks the candidate genes thought to synthesize the biomarker in question, it increases the probability that the biomarker in question is spurious. With awareness and vigilance, geochemists and molecular biologists can work together to identify and interpret authentic biomarker signatures.

The genomic record also has its limitations. Despite the exponential increase in genetic data, many clades most relevant to exotic biomarker production remain poorly represented. An example of this problem comes from the 'sponge' biomarker 24-isopropylcholesterol. Although there is evidence that gene duplication events allow sponges to synthesize exotic (C_{30+}) sterols, there is insufficient genetic coverage to determine when and how many times such duplication events have occurred. This is crucially important for interpreting whether putative sponge biomarkers indicate primitive proto-sponges or representatives of advanced modern groups. In addition to limited taxon sampling, there is the ever-present risk of unknown enzymes having a role in biomarker biosynthesis or unappreciated processes affecting their preservation. New enzymes are being discovered, even in well-studied biosynthesis pathways; examples include sterol demethylase proteins that allow bacteria to demethylate sterols[98,133], and the AltSQE enzyme that allows eukaryotes to oxidize squalene when they lack a canonical SQMO protein[98,134].

An exciting conceivable future application of lipid biomarkers is understanding potential extraterrestrial microbial life. Hydrocarbons, because of their relative recalcitrance and resistance to chemical, hydrolytic and thermal degradation, have significant potential to serve as biomarkers for extraterrestrial life. Indeed, a form of life alien to that found on the Earth would still be recognizable as 'life' if it produced organic compounds that exhibited characteristics consistent with systematic construction as opposed to the 'randomness' evident in, say, the molecules that have been identified in carbonaceous chondrites[135]. The molecules of life, as we know it, are constructed from a limited set of universal precursors — the 20 amino acids of protein, the 4 nucleobases of DNA and the 2 building blocks of lipids, which are acetate and isoprene. The universality of terrestrial biochemistry, therefore, results in assemblages of molecules that exhibit patterns in their atomic ratios and chemical structures, constitutional isomers that are limited in number and with isotopic compositions that are systematic in nature[136,137]. This is a quite distinct feature of fossil molecules preserved in ancient sediments on the Earth and there is no reason to expect otherwise on another planetary body. Despite the fact that molecular fossils become altered over time, the patterning encoded by universal biosynthetic pathways is a recognizable biosignature even for biochemistries that differ from those on the Earth[138,139].

As with other sciences, interpretation of geochemical biomarkers is continually undergoing refinement or revision. Some hypotheses — such as those based on the TEX_{86} SST proxy — rely on the principle of uniformitarianism, which may not fully extend to deep timescales. Similarly, there is always the possibility that exotic geological biomarkers were produced by a now-extinct life form; a risk that increases the farther back in time we go. But identifying the genes that underlie biomarker production and contextualizing them in a phylogenetic framework is a means to test competing hypotheses and refute inaccurate interpretations. The combination of genomics and geology promises to substantially advance our understanding of the Earth's chemical fossil record and, by extension, the nature of the ancient microbial life. This knowledge will grow along with the databases of organismic and environmental nucleic acids and improvements in the ways that this information can be queried and applied.

Published online 11 October 2021

1. Berner, E. K. & Berner, R. A. *Global Environment: Water, Air, and Geochemical Cycles* (Princeton Univ. Press, 2012).
2. Cavosie, A. J., Valley, J. W. & Wilde, S. A. The oldest terrestrial mineral record: a review of 4400 to 4000 Ma detrital zircons from Jack Hills, Western Australia. *Dev. Precambrian Geol.* **15**, 91–111 (2007).
3. Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
4. McNaughton, N. J., Compston, W. & Barley, M. E. Constraints on the age of the Warrawoona Group, eastern Pilbara Block, Western Australia. *Precambrian Res.* **60**, 69–98 (1993).
5. Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K. & Takeuchi, M. Microfossil assemblage from the 3400 Ma strelley pool formation in the Pilbara Craton, Western Australia: results form a new locality. *Precambrian Res.* **226**, 59–74 (2013).
6. Sugitani, K. et al. Early evolution of large microorganisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology* **13**, 507–521 (2015).
7. Alleon, J. et al. Chemical nature of the 3.4 Ga Strelley Pool microfossils. *Geochem. Perspect. Lett.* **7**, 37–42 (2018).
8. Allwood, A. C. et al. Controls on development and diversity of Early Archean stromatolites. *Proc. Natl Acad. Sci. USA* **106**, 9548–9555 (2009).
9. Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P. & Burch, I. W. Stromatolite reef from the Early Archaean era of Australia. *Nature* **441**, 714–718 (2006).
This paper details connections between the morphology of some of the oldest stromatolites and features of their coastal marine setting. It is key to illustrating how complex microbial communities must have existed on the Earth at least 3.45 billion years ago.
10. Hofmann, H., Grey, K., Hickman, A. & Thorpe, R. Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia. *Geol. Soc. Am. Bull.* **111**, 1256–1262 (1999).
11. Des Marais, D. J. Isotopic evolution of the biogeochemical carbon cycle during the Precambrian. *Rev. Mineral. Geochem.* **43**, 555–578 (2001).
12. Buick, R. et al. Record of emergent continental crust ~3.5 billion years ago in the Pilbara Craton of Australia. *Nature* **375**, 574–577 (1995).
13. Ueno, Y., Ono, S., Rumble, D. & Maruyama, S. Quadruple sulfur isotope analysis of ca. 3.5 Ga dresser formation: new evidence for microbial sulfate reduction in the early Archean. *Geochim. Cosmochim. Acta* **72**, 5675–5691 (2008).
14. Bontognali, T. R. R. et al. Sulfur isotopes of organic matter preserved in 3.45-billion-year-old stromatolites reveal microbial metabolism. *Proc. Natl Acad. Sci. USA* **109**, 15146–15151 (2012).
15. Beaumont, V. & Robert, F. Nitrogen isotope ratios of kerogens in Precambrian cherts: a record of the evolution of atmosphere chemistry? *Precambrian Res.* **96**, 63–82 (1999).
16. Morgan, G. J. Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959–1965. *J. Hist. Biol.* **31**, 155–178 (1998).

17. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965). **This classic paper informs us how the sequences of present-day macromolecules encode a history of their origin and evolution**.
18. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* 97–166 (Elsevier, 1965).
19. Peterson, K. J., Summons, R. E. & Donoghue, P. C. J. Molecular palaeobiology. *Palaeontology* **50**, 775–809 (2007).
20. Gaucher, E. A. Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zuckerkandl and Pauling. *Liberles [83]* **31**, 20–33 (2007).
21. Kacar, B., Hanson-Smith, V., Adam, Z. R. & Boekelheide, N. Constraining the timing of the Great Oxidation Event within the Rubisco phylogenetic tree. *Geobiology* **15**, 628–640 (2017).
22. Brocks, J. J. et al. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437**, 866–870 (2005).
23. McKenna, E. J. & Kallio, R. E. Microbial metabolism of the isoprenoid alkane pristane. *Proc. Natl Acad. Sci. USA* **68**, 1552 (1971).
24. Waples, D. W., Haug, P. & Welte, D. H. Occurrence of a regular C_{25} isoprenoid hydrocarbon in Tertiary sediments representing a lagoonal-type, saline environment. *Geochim. Cosmochim. Acta* **38**, 381–387 (1974).
25. Knoll, A. H., Summons, R. E., Waldbauer, J. R. & Zumberge, J. in *The Evolution of Primary Producers in the Sea* (eds Falkwoski, P. & Knoll, A.H.) 133–163 (Elsevier, 2007).
26. Brocks, J. J. The transition from a cyanobacterial to algal world and the emergence of animals. *Emerg. Top. Life Sci.* **2**, 181–190 (2018).
27. Sinninghe Damsté, J. S. & Köster, J. A euxinic southern North Atlantic Ocean during the Cenomanian/Turonian oceanic anoxic event. *Earth Planet. Sci. Lett.* **158**, 165–173 (1998).
28. Kuypers, M. M. M. et al. Massive expansion of marine archaea during a mid-cretaceous oceanic anoxic event. *Science* **293**, 92–95 (2001).
29. Brassell, S. C., Eglinton, G., Marlowe, I. T., Pflaumann, U. & Sarnthein, M. Molecular stratigraphy: a new tool for climatic assessment. *Nature* **320**, 129–133 (1986). **This study is the first detailing how fossilized organic molecules can serve as SST proxies**.
30. Schouten, S. et al. Extremely high sea-surface temperatures at low latitudes during the Middle Cretaceous as revealed by archaeal membrane lipids. *Geology* **31**, 1069–1072 (2003).
31. Bobrovskiy, I., Hope, J. M., Krasnova, A., Ivantsov, A. & Brocks, J. J. Molecular fossils from organically preserved Ediacara biota reveal cyanobacterial origin for *Beltanelliformis*. *Nat. Ecol. Evol.* **2**, 437 (2018).
32. Evitt, W. R. A discussion and proposals concerning fossil dinoflagellates, hystrichospheres, and acritarchs, II. *Proc. Natl Acad. Sci. USA* **49**, 298 (1963).
33. Treibs, A. Chlorophyll- und Häminderivate in organischen Mineralstoffen [German]. *Angew. Chem.* **49**, 682–686 (1936).
34. Hills, I. R. & Whitehead, E. V. Triterpanes in optically active petroleum distillates. *Nature* **209**, 977–979 (1966).
35. Blumer, M. Pigments of a fossil echinoderm. *Nature* **188**, 1100–1101 (1960).
36. Ourisson, G., Albrecht, P. & Rohmer, M. The hopanoids. Palaeochemistry and biochemistry of a group of natural products. *Pure Appl. Chem.* **51**, 709–729 (1979). **This review details how a particular group of bacterial membrane lipids gave rise to a ubiquitous and abundant class of chemical fossils**.
37. Rohmer, M. & Ourisson, G. Dérivés du bactériohopane: variations structurales et répartition [French]. *Tetrahedron Lett.* **17**, 3637–3640 (1976).
38. Yon, D. A., Maxwell, J. R. & Ryback, G. 2,6,10-Trimethyl-7-(3-methylbutyl)-dodecane, a novel sedimentary biological marker compound. *Tetrahedron Lett.* **23**, 2143–2146 (1982).
39. Barrick, R. C., Hedges, J. I. & Peterson, M. L. Hydrocarbon geochemistry of the Puget Sound region — I. Sedimentary acyclic hydrocarbons. *Geochim. Cosmochim. Acta* **44**, 1349–1362 (1980).
40. Requejo, A. G. & Quinn, J. G. Geochemistry of C_{25} and C_{30} biogenic alkenes in sediments of the Narragansett Bay estuary. *Geochim. Cosmochim. Acta* **47**, 1075–1090 (1983).
41. Dunlop, R. W. & Jefferies, P. R. Hydrocarbons of the hypersaline basins of Shark Bay, Western Australia. *Org. Geochem.* **8**, 313–320 (1985).
42. Volkman, J. K., Barrett, S. M. & Dunstan, G. A. C_{25} and C_{30} highly branched isoprenoid alkenes in laboratory cultures of two marine diatoms.*Org. Geochem.* **21**, 407–414 (1994).
43. Sinninghe Damste, J. S. et al. The rise of the rhizosolenid diatoms. *Science* **304**, 584–587 (2004).
44. Rowland, S. J. et al. Factors influencing the distributions of polyunsaturated terpenoids in the diatom, *Rhizosolenia setigera*. *Phytochemistry* **58**, 717–728 (2001).
45. Blumer, M., Guillard, R. R. L. & Chase, T. Hydrocarbons of marine phytoplankton. *Mar. Biol.* **8**, 183–189 (1971).
46. Eglinton, G. & Hamilton, R. J. Leaf epicuticular waxes. *Science* **156**, 1322–1335 (1967).
47. Rohmer, M., Bouvier-Nave, P. & Ourisson, G. Distribution of hopanoid triterpanes in prokaryotes. *J. Gen. Microbiol.* **130**, 1137–1150 (1984).
48. Volkman, J. K. et al. Microalgal biomarkers: a review of recent research developments. *Org. Geochem.* **29**, 1163–1179 (1998). **This paper reviews the laborious but essential work of surveying biomarkers across living organisms. The distribution of biomarkers in modern algae provides a solid foundation on which molecular fossils have historically been interpreted**.
49. Sturt, H. F., Summons, R. E., Smith, K., Elvert, M. & Hinrichs, K.-U. Intact polar membrane lipids in prokaryotes and sediments deciphered by high-performance liquid chromatography/electrospray ionization multistage mass spectrometry — new biomarkers for biogeochemistry and microbial ecology. *Rapid Commun. Mass. Spectrom.* **18**, 617–628 (2004).
50. White, D. C. & Ringelberg, D. B. in *Techniques in Microbial Ecology*. (eds Burlage, R. S. et al.) 255–272 (Oxford Univ. Press, 1998).
51. Vestal, J. R. & White, D. C. Lipid analysis in microbial ecology. *Bioscience* **39**, 535–541 (1989).
52. Lipp, J. S. & Hinrichs, K.-U. Structural diversity and fate of intact polar lipids in marine sediments. *Geochim. Cosmochim. Acta* **73**, 6816–6833 (2009).
53. Rossel, P. E. et al. Intact polar lipids of anaerobic methanotrophic archaea and associated bacteria. *Org. Geochem.* **39**, 992–999 (2008).
54. Taylor, J. & Parkes, R. J. The cellular fatty acids of the sulphate-reducing bacteria, *Desulfobacter* sp., *Desulfobulbus* sp. and *Desulfovibrio desulfuricans*. *J. Gen. Microbiol.* **129**, 3303–3309 (1983).
55. Brocks, J. J. & Pearson, A. Building the biomarker tree of life. *Rev. Mineral. Geochem.* **59**, 233–258 (2005).
56. Volkman, J. K. Sterols and other triterpenoids: source specificity and evolution of biosynthetic pathways. *Org. Geochem.* **36**, 139–159 (2005).
57. Schouten, S., Hopmans, E. C. & Sinninghe Damsté, J. S. The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: a review. *Org. Geochem.* **54**, 19–61 (2013).
58. Peters, K. E., Walters, C. C. & Moldowan, J. M. *The Biomarker Guide* 2nd edn (Cambridge Univ. Press, 2005).
59. Pearson, A. 12.11 Lipidomics for geochemistry. *Treatise Geochem.* **12**, 291–336 (2014).
60. Newman, D. K., Neubauer, C., Ricci, J. N., Wu, C.-H. & Pearson, A. Cellular and molecular biological approaches to interpreting ancient biomarkers. *Annu. Rev. Earth Planet. Sci.* **44**, 493–522 (2016). **This paper details our changing understanding on the role of 2-methylhopanoids in bacteria, and how this change impacts our interpretation of the related molecular fossil. It provides a case study on the importance of knowing what a biomarker biologically does in a microbe, not just its presence or absence**.
61. Ochs, D., Kaletta, C., Entian, K. D., Beck-Sickinger, A. & Poralla, K. Cloning, expression, and sequencing of squalene-hopene cyclase, a key enzyme in triterpenoid metabolism. *J. Bacteriol.* **174**, 298–302 (1992).
62. Schmerk, C. L. et al. Elucidation of the *Burkholderia cenocepacia* hopanoid biosynthesis pathway uncovers functions for conserved proteins in hopanoid-producing bacteria. *Environ. Microbiol.* **17**, 735–750 (2015).
63. Welander, P. V. et al. Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology* **10**, 163–177 (2012).
64. Pearson, A., Flood Page, S. R., Jorgenson, T. L., Fischer, W. W. & Higgins, M. B. Novel hopanoid cyclases from the environment. *Environ. Microbiol.* **9**, 2175–2188 (2007). **This paper is the first example of using a biomarker biosynthesis gene, the squalene–hopene cyclase gene necessary for hopanoid production, to demonstrate the potential diversity of biomarker producers in environmental metagenomic data sets**.
65. Villanueva, L., Rijpstra, W. I. C., Schouten, S. & Damsté, J. S. S. Genetic biomarkers of the sterol–biosynthetic pathway in microalgae. *Environ. Microbiol. Rep.* **6**, 35–44 (2014).
66. Villanueva, L., Schouten, S. & Sinninghe Damsté, J. S. Depth-related distribution of a key gene of the tetraether lipid biosynthetic pathway in marine Thaumarchaeota. *Environ. Microbiol.* **17**, 3527–3539 (2015).
67. Banta, A. B., Wei, J. H. & Welander, P. V. A distinct pathway for tetrahymanol synthesis in bacteria. *Proc. Natl Acad. Sci. USA* **112**, 13478–13483 (2015).
68. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
69. Eglinton, G. & Calvin, M. Chemical fossils. *Sci. Am.* **216**, 32–43 (1967).
70. Jensen, S. V. L. Bacterial carotenoids. *Acta Chem. Scand.* **19**, 1025–30 (1965).
71. Jensen, S. V. L. Bacterial carotenoids XXII. *Acta Chem. Scand.* **21**, 2578–80 (1967).
72. Summons, R. E. & Powell, T. G. *Chlorobiaceae* in Paleozoic seas revealed by biological markers, isotopes and geology. *Nature* **319**, 763–765 (1986).
73. Abella, C., Montesinos, E. & Guerrero, R. in *Shallow Lakes Contributions to Their Limnology* 173–181 (Springer, 1980).
74. French, K. L., Rocher, D., Zumberge, J. E. & Summons, R. E. Assessing the distribution of sedimentary C_{40} carotenoids through time. *Geobiology* **13**, 139–151 (2015).
75. Sinninghe Damsté, J. S. & Koopmans, M. P. The fate of carotenoids in sediments: an overview. *Pure Appl. Chem.* **69**, 2067–2074 (1997).
76. Frigaard, N.-U., Maresca, J. A., Yunker, C. E., Jones, A. D. & Bryant, D. A. Genetic manipulation of carotenoid biosynthesis in the green sulfur bacterium *Chlorobium tepidum*. *J. Bacteriol.* **186**, 5210–5220 (2004).
77. Maresca, J., Graham, J. & Bryant, D. The biochemical basis for structural diversity in the carotenoids of chlorophototrophic bacteria. *Photosynthesis Res.* **97**, 121–140 (2008).
78. Maresca, J. A., Romberger, S. P. & Bryant, D. A. Isorenieratene biosynthesis in green sulfur bacteria requires the cooperative actions of two carotenoid cyclases. *J. Bacteriol.* **190**, 6384–6391 (2008).
79. Vogl, K. & Bryant, D. A. Biosynthesis of the biomarker okenone: χ-ring formation. *Geobiology* **10**, 205–215 (2012).
80. Krügel, H., Krubasik, P., Weber, K., Saluz, H. P. & Sandmann, G. Functional analysis of genes from *Streptomyces griseus* involved in the synthesis of isorenieratene, a carotenoid with aromatic end groups, revealed a novel type of carotenoid desaturase. *Biochim. Biophys. Acta* **1439**, 57–64 (1999).
81. Krubasik, P. & Sandmann, G. A carotenogenic gene cluster from *Brevibacterium linens* with novel lycopene cyclase genes involved in the synthesis of aromatic carotenoids. *Mol. Gen. Genet.* **263**, 423–432 (2000).
82. Graham, J. E., Lecomte, J. T. J. & Bryant, D. A. Synechoxanthin, an aromatic C_{40} xanthophyll that is a major carotenoid in the cyanobacterium *Synechococcus* sp. PCC 7002. *J. Nat. Products* **71**, 1647–1650 (2008).
83. Graham, J. E. & Bryant, D. A. The biosynthetic pathway for synechoxanthin, an aromatic carotenoid synthesized by the euryhaline, unicellular cyanobacterium *Synechococcus* sp. strain PCC 7002. *J. Bacteriol.* **190**, 7966–7974 (2008).
84. Koopmans, M. P., Schouten, S., Kohnen, M. E. L. & Damsté, J. S. S. Restricted utility of aryl isoprenoids as indicators for photic zone anoxia. *Geochim. Cosmochim. Acta* **60**, 4873–4876 (1996).
85. Brocks, J. J. & Schaeffer, P. Okenane, a biomarker for purple sulfur bacteria (Chromatiaceae), and other new carotenoid derivatives from the 1640 Ma Barney Creek formation. *Geochim. Cosmochim. Acta* **72**, 1396–1414 (2008).
86. Yamaguchi, M. On carotenoids of a sponge "*Reniera japonica*". *Bull. Chem. Soc. Jpn.* **30**, 111–114 (1957).
87. Yamaguchi, M. Renieratene, a new carotenoid containing benzene rings, isolated from a sea sponge. *Bull. Chem. Soc. Jpn.* **31**, 739–742 (1958).
88. Hentschel, U., Piel, J., Degnan, S. M. & Taylor, M. W. Genomic insights into the marine sponge microbiome. *Nat. Rev. Microbiol.* **10**, 641–654 (2012).
89. French, K. L., Birdwell, J. E. & Berg, V. Biomarker similarities between the saline lacustrine eocene green river and the paleoproterozoic Barney Creek formations. *Geochim. Cosmochim. Acta* **274**, 228–245 (2020).

90. Cui, X. et al. Niche expansion for phototrophic sulfur bacteria at the Proterozoic–Phanerozoic transition. *Proc. Natl Acad. Sci. USA* **117**, 17599–17606 (2020).
91. Koopmans, M. P., De Leeuw, J. W. & Sinninghe Damsté, J. S. Novel cyclised and aromatised diagenetic products of β-carotene in the Green River Shale. *Org. Geochem.* **26**, 451–466 (1997).
92. Behrens, A., Schaeffer, P., Bernasconi, S. & Albrecht, P. Mono- and bicyclic squalene derivatives as potential proxies for anaerobic photosynthesis in lacustrine sulfur-rich sediments. *Geochim. Cosmochim. Acta* **64**, 3327–3336 (2000).
93. Schaeffer, P., Adam, P., Wehrung, P. & Albrecht, P. Novel aromatic carotenoid derivatives from sulfur photosynthetic bacteria in sediments. *Tetrahedron Lett.* **38**, 8413–8416 (1997).
94. Brocks, J. J. et al. The rise of algae in Cryogenian oceans and the emergence of animals. *Nature* **548**, 578 (2017).
This study highlights how specific chemical modifications in lipid structures, in this case methylation of sterol molecules, can be informative and can be used to track the emergence of specific microbial groups in the geologic record.
95. Javaux, E. J. & Knoll, A. H. Micropalaeontology of the lower Mesoproterozoic Roper Group, Australia, and implications for early eukaryotic evolution. *J. Paleontol.* **91**, 199–229 (2017).
96. Knoll, A. H. The early evolution of eukaryotes: a geological perspective. *Science* **256**, 622–627 (1992).
97. Wei, J. H., Yin, X. & Welander, P. V. Sterol synthesis in diverse bacteria. *Front. Microbiol.* **7**, 990 (2016).
98. Hoshino, Y. & Gaucher, E. A. Evolution of bacterial steroid biosynthesis and its impact on eukaryogenesis. *Proc. Natl Acad. Sci. USA* **118**, e2101276118 (2021).
This recent study uses a phylogenetic approach to assess the evolutionary history of sterol biosynthesis and the potential impact of bacterial sterol biosynthesis on the rise of eukaryotes.
99. Holland, H. D. The oxygenation of the atmosphere and oceans. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 903–915 (2006).
100. Luo, G. et al. Rapid oxidation of Earth's atmosphere 2.33 billion years ago. *Sci. Adv.* **2**, e1600134 (2016).
101. Gold, D. A., Caron, A., Fournier, G. P. & Summons, R. E. Paleoproterozoic sterol biosynthesis and the rise of oxygen. *Nature* **543**, 420–423 (2017).
102. Barker, H. A. Studies upon the methane-producing bacteria. *Arch. für Mikrobiologie* **7**, 420–438 (1936).
103. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
This classic study shows how ribosomal RNA sequences reveal that all life follows one of three lines of descent from a common ancestor.
104. Spang, A., Caceres, E. F. & Ettema, T. J. G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).
105. Blank, C. E. Not so old archaea — the antiquity of biogeochemical processes in the archaeal domain of life. *Geobiology* **7**, 495–514 (2009).
106. Salvador-Castell, M., Tourte, M. & Oger, P. M. In search for the membrane regulators of archaea. *Int. J. Mol. Sci.* **20**, 4434 (2019).
107. Koga, Y. & Morii, H. Recent advances in structural research on ether lipids from archaea including comparative and physiological aspects. *Biosci. Biotechnol. Biochem.* **69**, 2019–2034 (2005).
108. Moldowan, J. M. & Seifert, W. K. Head-to-head linked isoprenoid hydrocarbons in petroleum. *Science* **204**, 169–171 (1979).
109. Baumann, L. M. F. et al. Intact polar lipid and core lipid inventory of the hydrothermal vent methanogens *Methanocaldococcus villosus* and *Methanothermococcus okinawensis*. *Org. Geochem.* **126**, 33–42 (2018).
110. Summons, R. E., Powell, T. G. & Boreham, C. J. Petroleum geology and geochemistry of the Middle Proterozoic McArthur Basin, northern Australia: III. Composition of extractable hydrocarbons. *Geochim. Cosmochim. Acta* **52**, 1747–1763 (1988).
111. Tierney, J. E. in *Treatise on Geochemistry* Vol. 12 (eds Holland, H.D. & Turekian, K.K.) 379–393 (Elsevier, 2014).
112. Weijers, J. W. H., Schouten, S., van den Donker, J. C., Hopmans, E. C. & Sinninghe Damsté, J. S. Environmental controls on bacterial tetraether membrane lipid distribution in soils. *Geochim. Cosmochim. Acta* **71**, 703–713 (2007).
113. Schouten, S., Forster, A., Panoto, F. E. & Sinninghe Damsté, J. S. Towards calibration of the TEX_8^6 palaeothermometer for tropical sea surface temperatures in ancient greenhouse worlds. *Org. Geochem.* **38**, 1537–1546 (2007).
114. Schouten, S., Hopmans, E. C., Schefuβ, E. & Sinninghe Damsté, J. S. Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? *Earth Planet. Sci. Lett.* **204**, 265–274 (2002).
This study establishes the basis for the TEX_{86} palaeotemperature proxy as a SST based on the distribution of archaeal GDGT membrane lipids in marine sediments.
115. Schouten, S., Hopmans, E. C. & Damsté, J. S. S. The effect of maturity and depositional redox conditions on archaeal tetraether lipid palaeothermometry. *Org. Geochem.* **35**, 567–571 (2004).
116. Tierney, J. E. GDGT thermometry: lipid tools for reconstructing paleotemperatures. *Paleontol. Soc. Pap.* **18**, 115–132 (2012).
117. Zhang, Y. G., Pagani, M. & Wang, Z. Ring Index: a new strategy to evaluate the integrity of TEX_{86} paleothermometry. *Paleoceanography* **31**, 220–232 (2016).
118. Kim, J.-H., Schouten, S., Hopmans, E. C., Donner, B. & Sinninghe Damsté, J. S. Global sediment core-top calibration of the TEX_{86} paleothermometer in the ocean. *Geochim. Cosmochim. Acta* **72**, 1154–1173 (2008).
119. Kim, J.-H. et al. New indices and calibrations derived from the distribution of crenarchaeal isoprenoid tetraether lipids: implications for past sea surface temperature reconstructions. *Geochim. Cosmochim. Acta* **74**, 4639–4654 (2010).
120. Trommer, G. et al. Distribution of Crenarchaeota tetraether membrane lipids in surface sediments from the Red Sea. *Org. Geochem.* **40**, 724–731 (2009).
121. Tierney, J. E. & Tingley, M. P. A Bayesian, spatially-varying calibration model for the TEX_{86} proxy. *Geochim. Cosmochim. Acta* **127**, 83–106 (2014).
122. Tierney, J. E. & Tingley, M. P. A TEX_{86} surface sediment database and extended Bayesian calibration. *Sci. Data* **2**, 150029 (2015).
123. Zhou, A. et al. Energy flux controls tetraether lipid cyclization in *Sulfolobus acidocaldarius*. *Environ. Microbiol.* **22**, 343–353 (2020).
124. Qin, W. et al. Confounding effects of oxygen and temperature on the TEX_{86} signature of marine Thaumarchaeota. *Proc. Natl Acad. Sci. USA* **112**, 10979–10984 (2015).
125. Hurley, S. J. et al. Influence of ammonia oxidation rate on thaumarchaeal lipid composition and the TEX_{86} temperature proxy. *Proc. Natl Acad. Sci. USA* **113**, 7762–7767 (2016).
126. DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl Acad. Sci. USA* **89**, 5685–5689 (1992).
127. Lincoln, S. A. et al. Planktonic Euryarchaeota are a significant source of archaeal tetraether lipids in the ocean. *Proc. Natl Acad. Sci. USA* **111**, 9858–9863 (2014).
128. Zeng, Z. et al. GDGT cyclization proteins identify the dominant archaeal sources of tetraether lipids in the ocean. *Proc. Natl Acad. Sci. USA* **116**, 22505–22511 (2019).
129. Besseling, M. A. et al. The absence of intact polar lipid-derived GDGTs in marine waters dominated by Marine Group II: implications for lipid biosynthesis in archaea. *Sci. Rep.* **10**, 1–10 (2020).
130. Pearson, A. Resolving a piece of the archaeal lipid puzzle. *Proc. Natl Acad. Sci. USA* **116**, 22423–22425 (2019).
131. Gold, D. A., O'Reilly, S. S., Luo, G., Briggs, D. E. G. & Summons, R. E. Prospects for sterane preservation in sponge fossils from museum collections and the utility of sponge biomarkers for molecular clocks. *Bull. Peabody Mus. Nat. History* **57**, 181–189 (2016).
132. French, K. L. et al. Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proc. Natl Acad. Sci. USA* **112**, 5915–5920 (2015).
133. Lee, A. K. et al. C-4 sterol demethylation enzymes distinguish bacterial and eukaryotic sterol synthesis. *Proc. Natl Acad. Sci. USA* **115**, 5884–5889 (2018).
134. Pollier, J. et al. A widespread alternative squalene epoxidase participates in eukaryote steroid biosynthesis. *Nat. Microbiol.* **4**, 226–233 (2019).
135. Cronin, J. R., Pizzarello, S., Epstein, S. & Krishnamurthy, R. V. Molecular and isotopic analyses of the hydroxy acids, dicarboxylic acids, and hydroxydicarboxylic acids of the Murchison meteorite. *Geochim. Cosmochim. Acta* **57**, 4745–4752 (1993).
136. Summons, R. E., Albrecht, P., McDonald, G. & Moldowan, J. M. Molecular biosignatures. *Strateg. Life Detection* **25**, 133–159 (2008).
137. Davila, A. F. & McKay, C. P. Chance and necessity in biochemistry: implications for the search for extraterrestrial biomarkers in Earth-like environments. *Astrobiology* **14**, 534–540 (2014).
138. Summons, R. E. et al. Preservation of martian organic and environmental records: final report of the Mars Biosignature Working Group. *Astrobiology* **11**, 157–181 (2011).
139. McKay, C. P. What is life — and how do we search for it in other worlds? *PLoS Biol.* **2**, e302 (2004).
140. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth/'s early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
141. Martin, A. P., Condon, D. J., Prave, A. R. & Lepland, A. A review of temporal constraints for the Palaeoproterozoic large, positive carbonate carbon isotope excursion (the Lomagundi–Jatuli Event). *Earth Sci. Rev.* **127**, 242–261 (2013).
142. Welander, P. V., Coleman, M., Sessions, A. L., Summons, R. E. & Newman, D. K. 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 (2010).
143. Zundel, M. & Rohmer, M. Prokaryotic triterpenoids. 3. The biosynthesis of 2β-methylhopanoids and 3β-methylhopanoids of *Methylobacterium organophilum* and *Acetobacter pasteurianus* ssp. *pasteurianus*. *Eur. J. Biochem.* **150**, 35–39 (1985).
144. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
145. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
146. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
147. Schmerk, C. L., Bernards, M. A. & Valvano, M. A. Hopanoid production is required for low-pH tolerance, antimicrobial resistance, and motility in *Burkholderia cenocepacia*. *J. Bacteriol.* **193**, 6712–6723 (2011).
148. Ricci, J. N., Morton, R., Kulkarni, G., Summers, M. L. & Newman, D. K. Hopanoids play a role in stress tolerance and nutrient storage in the cyanobacterium *Nostoc punctiforme*. *Geobiology* **15**, 173–183 (2017).
149. Garby, T. J. et al. Lack of methylated hopanoids renders the cyanobacterium *Nostoc punctiforme* sensitive to osmotic and pH stress. *Appl. Environ. Microbiol.* **83**, e00777–00717 (2017).
150. Bradley, A. S. et al. Hopanoid-free *Methylobacterium extorquens* DM4 overproduces carotenoids and has widespread growth impairment. *PLoS ONE* **12**, e0173323 (2017).
151. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).
152. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).
153. Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: statistically informed gene tree error correction using species trees. *Syst. Biol.* **62**, 110–120 (2013).
154. Magnabosco, C., Moore, K. R., Wolfe, J. M. & Fournier, G. P. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology* **16**, 179–189 (2018).
155. Brasier, M. D. et al. Questioning the evidence for Earth's oldest fossils. *Nature* **416**, 76–81 (2002).
156. Knoll, A. H., Bergmann, K. D. & Strauss, J. V. Life: the first two billion years. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150493 (2016).

Acknowledgements
This work was supported by a grant from the Simons Collaboration on the Origin of Life through an award (290361FY18) to R.E.S. and by National Science Foundation awards (1752564) to P.V.W and (2044871) to D.A.G.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Microbiology thanks V. Parro García, J. Peckmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.