Paleoproterozoic sterol biosynthesis and the rise of oxygen

Citation: Gold, David A. et al. "Paleoproterozoic sterol biosynthesis and the rise of oxygen." Nature 543, 7645 (March 2017): 420–423. © 2017 Macmillan Publishers Limited, part of Springer Nature

As Published: http://dx.doi.org/10.1038/nature21412

Publisher: Springer Science and Business Media LLC

Persistent URL: https://hdl.handle.net/1721.1/128450

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Paleoproterozoic sterol biosynthesis and the rise of oxygen

[1, 2] David A. Gold (dagold@mit.edu)

[1]Abigail Caron (abigailc@mit.edu)

[1]Gregory P. Fournier (g4nier@mit.edu)

[1]Roger E. Summons* (rsummons@mit.edu)

1. Department of Earth, Atmospheric and Planetary Science, Massachusetts Institute of Technology, Cambridge, MA, USA.
2. Present Address: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA.

* Corresponding author: rsummons@mit.edu

Text

Natural products preserved in the geologic record can function as "molecular fossils", providing insight into organisms and physiologies that existed in the deep past. One important group of molecular fossils is the steroidal hydrocarbons (steranes), which are the diagenetic remains of sterol lipids. Complex sterols with modified side chains are unique to eukaryotes, although simpler sterols can also be synthesized by a few bacteria [1]. Sterol biosynthesis is an oxygen-intensive process; thus, the presence of complex steranes in ancient rocks not only signals the presence of eukaryotes, but also aerobic metabolic processes [2]. In 1999, steranes were reported in 2.7 billion year old (Gyr-old) rocks from the Pilbara Craton in Australia [3], suggesting a long delay between photosynthetic oxygen production and its accumulation in the atmosphere (also known as the Great Oxidation Event) 2.45–2.32 Gyr ago [4]. However, the recent reappraisal and rejection of these steranes as contaminants [5] pushes the oldest reported steranes forward to ~1.64 Gyr [6]. In the present study, we use a molecular clock approach to improve constraints on the evolution of sterol biosynthesis. We infer that stem-eukaryotes shared functionally modern sterol biosynthesis genes with bacteria *via* horizontal gene transfer. Comparing multiple molecular clock analyses, we find that the maximum marginal probability for the divergence time of bacterial and eukaryal sterol biosynthesis genes is ~2.31 Gyr ago, concurrent with the most recent geochemical evidence for the Great Oxidation Event (GOE) [7]. Our results therefore suggest that simple sterol biosynthesis existed well before the diversification of living eukaryotes, substantially predating the oldest detected sterane biomarkers (~1.64 Gyr [6]), and furthermore, that the evolutionary history of sterol

biosynthesis is tied to the first widespread availability of molecular oxygen in the ocean-atmosphere system.

For this study we focused on the first two enzymes necessary for sterol biosynthesis: squalene monooxygenase (SQMO; enzyme commission number 1.14.14.17) and oxidosqualene cyclase (OSC; EC number 5.4.99.7). These proteins use molecular oxygen to sequentially convert squalene into the protosterol precursors required for complex eukaryotic sterols; [8] they are also the only two enzymes in the canonical sterol biosynthesis pathway (KEGG reference map00100) typically found in bacteria other than squalene synthetase, which is broadly conserved across all domains of life. We wrote a set of scripts (detailed in the Methods) to vet all potential SQMO and OSC proteins in the NCBI protein database, map their distribution across taxa, and subsample the data based on taxonomic completeness. To root these gene trees for phylogenetic analyses, we also collected data from relevant protein paralogs: rooting OSC with squalene hopene cyclase (SHC; EC number 5.4.99.17), and rooting SQMO with the ubiH/COQ6 family, which includes the eukaryote-specific ubiquinone biosynthesis monooxygenase (COQ6), as well as the bacterial ubiquinone biosynthesis hydroxylase family (UbiH/UbiF/VisC). This protocol ensured a complete collection of prokaryotic SQMO and OSC proteins in NCBI, as well as a well-sampled eukaryotic dataset.

Phylogenetic analyses of *sqmo* and *osc* suggest that the two genes have a shared evolutionary history (Fig. 1a). The eukaryotic portions of the *sqmo* and *osc* gene trees broadly recapitulate the most current species trees [9,10] albeit with low support at the

deepest nodes (see Extended Data Figs. 1-6). This supports the hypothesis that both genes were present in the last common ancestor of eukaryotes [8]. Some sequences from amoebozoan, alveolate, and rhizarian taxa have unconventional placements within these phylogenies; this is likely caused by long-branch attraction and other phylogenetic artifacts, but we cannot rule out limited horizontal gene transfer between eukaryotes (see Extended Data Figs. 1 and 2). We found no evidence for *sqmo* or *osc* genes in Archaea, but did confirm one or both genes in 27 bacterial taxa, representing 6 phyla, 9 classes, and 9 orders (Fig. 1b). In phylogenetic analyses, all bacterial sequences group in one of two places: either basal to crown-group eukaryotes ("Bacterial Group I"), or nested within the bikonts ("Bacterial Group II"). Similar trees were recently reported in a study of bacterial *osc* genes [1], but the consistency between *osc* and *sqmo* topologies has not previously been observed. The consistent grouping of bacterial taxa into either Group I or II suggests that *sqmo* and *osc* have moved together through horizontal gene transfer (Fig. 1a). The linked inheritance of these two genes is also supported by synteny analysis of bacterial genomes, with *sqmo* and *osc* rarely being separated by more than one gene (Fig. 1b). Such synteny has previously been observed in two species of bacteria [11], but this study suggests the phenomenon is broadly conserved. The only taxon where *sqmo* and *osc* demonstrate conflicting phylogenic placement is *Eudora adriatica*; because genomic synteny is conserved in this bacterium, the inconsistency is unlikely to represent independent gene transfers, and is most likely a long-branch artifact.

The distribution of *sqmo* and *osc* genes can be explained by at least two horizontal gene transfer events between bacteria and eukaryotes. The nesting of Bacterial Group II within

the crown-group bikonts strongly suggests a horizontal gene transfer from eukaryotes into bacteria. Conversely, Bacterial Group I roots outside of extant Eukarya, so the directionality of this horizontal gene transfer event cannot be unambiguously determined from our data by means of a polarizing outgroup. However, the shallow clade depth and sparse phylogenetic distribution of bacterial genes—combined with the relatively long branches leading to eukaryal representatives—are most consistent with these genes being transferred one or more times from stem-eukaryotes to bacteria [11]. Additionally, because Bacterial Group I includes genes from *Gemmata*—a genus that has been demonstrated to use SQMO and OSC enzymes to produce protosterols [11]—we can infer that these horizontal gene transfer events involved the sharing of functionally modern proteins, permitting the reconstruction of the character state for these genes at the coalescent node. In this way, horizontal gene transfer events can be used to map the presence of characters onto stem lineages in the absence of paleontological evidence.

Insert Figure 1 hereabouts

To study the timing of the Bacterial Group I / stem-eukaryote split, we performed a series of molecular clock analyses (Fig. 2 and Table 1). We tested SQMO and OSC separately—with and without sister genes as outgroups—as well as the two proteins concatenated into a single dataset. Our gene trees suggest that excavates are the earliest branching eukaryote lineage, a hypothesis that is currently controversial [9,10]. We therefore repeated all analyses using an alternative topology where excavates are sister to bikonts. The timing of the Bacterial Group I / stem-eukaryote split has large uncertainties associated with it—as expected given the limited data in single-gene analyses. The 95%

confidence intervals consistently fall between 1.75 and 3.05 Gyr, with one exception coming from our SQMO-only datasets, which give significantly older dates than other analyses (Table 1). But when SQMO is constrained by employing the UbiH/COQ6 family as an outgroup, it produces results congruent with the other analyses. The SQMO-only datasets also produce an origin for crown-group eukaryotes that is significantly older than estimates from previous multi-gene molecular clock analyses [12-14], suggesting that molecular clocks derived from SQMO-only data lead to a general overestimation of true divergence times.

Insert Figure 2 hereabouts

The marginal probabilities associated with the Bacterial Group I / stem-eukaryote split are greatest around the time period of the Great Oxidation Event (Fig. 3). The marginal probability curves in the SQMO-only datasets are significantly older than the others, but also have the lowest peak densities, again suggesting that these analyses should be viewed with caution. Using our preferred analysis (a concatenated SQMO and OSC sequences with an excavates-basal tree), we specifically recover a 94.5% probability of this younger-bound age constraint on oxygen-dependent sterol biosynthesis predating the Orosirian Period (>2.05 Ga; see Extended Table 1 for the distribution of marginal probabilities for all analyses across geologic time). If we treat excavates as sister to the bikonts (instead of sister to all eukaryotes, as our analysis supports), we recover slightly younger dates, but the change does not fundamentally impact our interpretation of the data. All analyses suggest a Paleoproterozoic-or-earlier existence for *sqmo* and *osc* genes.

Insert Figure 3 hereabouts

In conclusion, our molecular clock analyses suggest that protosterol biosynthesis likely existed by the time oxygen is first detectable as a permanent presence in the Earth's atmosphere [4,7,15,16]. Our results similarly suggest that sterol biosynthesis substantially predates the evolution of crown-group eukaryotes, and was likely an important preadaptation to modern eukaryotic life [17]. Finally, our results are inconsistent with an origin of sterol biosynthesis ~1.64 Gyr ago, as informed by the oldest sterane biomarkers currently known [6]. Reconciling the molecular clocks with the Proterozoic biomarker record requires considerable caution following the disproving of earlier reports of Archaean steranes [5], and the demonstration of widespread fossil fuel contamination in laboratory aerosols [18]. Most reliance can be placed on studies conducted on low maturity, organic-rich rock sequences and which have been replicated [6,19] and/or supported using multiple technologies [20]. The progression of sterols identified in such studies—from simple steroids at ~1.64 Gyr ago [6], to atypical triterpenoid biomarker patterns in the Neoproterozoic[19,21], to unambiguous and abundant eukaryotic sterols with modified side-chains in the Phanerozoic—is consistent with the genetic data, but molecular clocks suggest this progression must have occurred much earlier. Regarding Phanerozoic sterols with modified side-chains, the conservation of the sterol biosynthesis pathway across eukaryotes [8,22] means that such sterols were being synthesized by the ancestral crown-group eukaryote, which we date between ~1.30-2.17 Gyr ago, and multi-gene molecular clocks estimate between 0.95 and 1.87 Gyr ago [12-14]. The presence of atypical sterane

patterns in 0.8-1.64 Gyr-old rocks is therefore consistent with crown-group eukaryotic sources, although we cannot rule out non-eukaryotic interpretations for some of these biomarkers [19,21,23]. Regarding simple steroids, the most conservative interpretation of our data (i.e. the youngest date falling within a 95% confidence interval in any of our analyses) suggests that protosterols were being synthesized >1.75 Gyr ago. Using more realistic estimates (based on the averaged means of all analyses, excluding the SQMO-only datasets as outliers) we find that basic protosterols were likely being synthesized >2.31 Gyr ago. This suggests a >670 Myr gap between our age estimates and the oldest fossil steranes. This gap between the molecular and geochemical evidence for sterol biosynthesis, which may reflect sampling bias or could have ecological[24] or taphonomic explanations[5], will only be resolved by further discovery (see Supplemental Text for a more detailed discussion).

References

1 Wei, J. H., Yin, X. & Welander, P. V. Sterol Synthesis in Diverse Bacteria. *Frontiers in Microbiology* **7**, 990, doi:10.3389/fmicb.2016.00990 (2016).

2 Summons, R. E., Bradley, A. S., Jahnke, L. L. & Waldbauer, J. R. Steroids, triterpenoids and molecular oxygen. *Philosophical Transactions of the Royal Society B-Biological Sciences* **361**, 951-968, doi:10.1098/rstb.2006.1837|ISSN 0962-8436 (2006).

3 Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science* **285**, 1033-1036, doi:10.1126/science.1096806 (1999).

4 Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427**, 117-120 (2004).

5 French, K. L. *et al.* Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proceedings of the National Academy of Sciences* **112**, 5915-5920 (2015).

6 Brocks, J. J. *et al.* Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437**, 866 (2005).

7 Luo, G. *et al.* Rapid oxidation of Earth's atmosphere 2.33 billion years ago. *Science Advances* **2**, e1600134, doi:10.1126/sciadv.1600134 (2016).

8 Desmond, E. & Gribaldo, S. Phylogenomics of Sterol Synthesis: Insights into the Origin, Evolution, and Diversity of a Key Eukaryotic Feature. *Genome Biology and Evolution* **1**, 364-381, doi:10.1093/gbe/evp036 (2009).
9 He, D. *et al.* An Alternative Root for the Eukaryote Tree of Life. *Current Biology* **24**, 465-470, doi:http://dx.doi.org/10.1016/j.cub.2014.01.036 (2014).
10 Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E693-E699 (2015).
11 Pearson, A., Budin, M. & Brocks, J. J. Phylogenetic and biochemical evidence for sterol synthesis in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15352-15357 (2003).
12 Douzery, E. J. P., Snell, E. A., Bapteste, E., Delsuc, F. & Philippe, H. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15386-15391 (2004).
13 Berney, C. & Pawlowski, J. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proceedings of the Royal Society of London B: Biological Sciences* **273**, 1867-1872 (2006).
14 Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences* **108**, 13624-13629 (2011).
15 Farquhar, J., Bao, H. & Thiemens, M. Atmospheric Influence of Earth's Earliest Sulfur Cycle. *Science* **289**, 756-758, doi:10.1126/science.289.5480.756 (2000).
16 Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth/'s early ocean and atmosphere. *Nature* **506**, 307-315, doi:10.1038/nature13068 (2014).
17 Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *International Journal Of Systematic And Evolutionary Microbiology* **52**, 297-354 (2002).
18 Illing, C. J., Hallmann, C., Miller, K. E., Summons, R. E. & Strauss, H. Airborne hydrocarbon contamination from laboratory atmospheres. *Organic Geochemistry* **76**, 26-38, doi:http://dx.doi.org/10.1016/j.orggeochem.2014.07.006 (2014).
19 Summons, R. E. *et al.* Distinctive hydrocarbon biomarkers from fossiliferous sediment of the Late Proterozoic Walcott Member, Chuar Group, Grand Canyon, Arizona. *Geochimica et Cosmochimica Acta* **52**, 2625 (1988).
20 Love, G. D., Stalvies, C., Grosjean, E., Meredith, W. & Snape, C. E. Analysis of molecular biomarkers covalently bound within Neoproterozoic sedimentary kerogen in *Paleontological Society Papers* Vol. 14 (eds P.H. Kelley & R.K. Bambach) 67-83 (The Paleontological Society, 2008).
21 Brocks, J. J. *et al.* Early sponges and toxic protists: possible sources of cryostane, an age diagnostic biomarker antedating Sturtian Snowball Earth. *Geobiology* **14**, 129-149, doi:10.1111/gbi.12165 (2016).
22 Gold, D. A. *et al.* Sterol and genomic analyses validate the sponge biomarker hypothesis. *Proceedings of the National Academy of Sciences* **113**, 2684-2689 (2016).

23 Banta, A. B., Wei, J. H. & Welander, P. V. A distinct pathway for tetrahymanol synthesis in bacteria. *Proceedings of the National Academy of Sciences* **112**, 13478-13483 (2015).
24 Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: A bioinorganic bridge? *Science* **297**, 1137-1142 (2002).

Legends to Figures

Figure 1. Phylogeny and synteny of *sqmo* and *osc* genes. (A) Maximum likelihood trees for *SQMO* (left) and *OSC* (right) based on Extended Data Figs. 3 and 4. Note that the two trees provide consistent topologies, with bacteria clustering within crown-group eukaryotes ("Bacterial Group II") or outside ("Bacterial Group I"). Some eukaryotes have been excluded from these trees; see Extended Data Figs. 1 and 2. (B) Distribution of vetted bacterial *SQMO* and *OSC* genes, divided into the two phylogenetic groups. For species containing both genes, the relevant placement of both genes on the genome is provided to the left. Additional genes of interest have also been colored, and are described in the key. Note how there is greater evidence of synteny in Bacterial Group II, consistent with this clade representing a more recent horizontal gene transfer event.

Figure 2. Molecular clock for one of the datasets used in this study (SQMO + OSC concatenated together, with excavates as the earliest-branching eukaryotes). Bars represent 95% confidence intervals for nodes with >70% posterior probability. See Table 1 for the results of all analyses. Stars signify important nodes included in Table 1, circles indicate fossil calibrations.

Figure 3. Marginal probability curves for the timing of the Bacterial Group I / stem-eukaryote split (red star in Figure 2). Differences in clock estimation times are compared for our preferred topology (A), where excavates are the earliest-branching eukaryotes, and (B) where excavates are sister to the bikonts. Curves were generated by sampling the MCMC analysis every 1,000 generations for 20 million generations, with a 25% burn-in.

Acknowledgements

We gratefully acknowledge funding from the Agouron Institute Geobiology Fellowship to D.A.G. and the Simons Foundation Collaboration on the Origins of Life (SCOL) to R.E.S. Additional support was provided by the NSF program 'Frontiers of Earth System Dynamics' (EAR-1338810) to R.E.S.

Author contributions

R.E.S. and D.A.G. designed the experiment. D.A.G. and A.M.C. performed the data analysis. All authors were involved in interpreting the data and drafting the manuscript.

Competing financial interests.

The authors declare no competing financial interests.

Materials & Correspondence.

Correspondence to: Roger Summons (rsummons@mit.edu)

Table

Table 1. Mean values for important nodes in molecular clock analyses (in Gyr), with 95% confidence intervals in parentheses.

Gene	Gene outgroup included?	Excavates basal eukaryotes?	Origin of Bacterial Group I	Origin of crown-eukaryote SQMO/OSC	Origin of Bacterial Group II
SQMO	No	Yes	3.21(2.49-4.03)	2.22 (1.78-2.73)	1.54 (1.30-1.83)
SQMO	No	No	3.27 (2.51-4.18)	1.98 (1.58-2.34)	1.61 (1.31-1.92)
SQMO	Yes	Yes	2.47 (1.98-3.05)	1.80 (1.51-2.12)	1.35 (1.22-1.50)
SQMO	Yes	No	2.34(1.78-2.95)	1.60(1.40-1.92)	1.36(1.23-1.53)
OSC	No	Yes	2.23(1.78-2.77)	1.61(1.39-1.87)	1.34(1.15-1.57)
OSC	No	No	2.23(1.75-2.75)	1.51(1.30-1.81)	1.34(1.10-1.64)
OSC	Yes	Yes	2.34(1.93-2.82)	1.56(1.37-1.78)	1.32(1.14-1.51)
OSC	Yes	No	2.31(1.95-2.73)	1.51(1.33-1.70)	1.35(1.16-1.55)
Both	No	Yes	2.28(1.90-2.71)	1.82(1.56-2.17)	1.46(1.28-1.65)
Both	No	No	2.30(1.87-2.78)	1.60(1.45-1.80)	1.44(1.30-1.61)

Methods

Data availability. All amino acid alignments and trees from this study are available as a Supplementary Information file. GI numbers for sequences used in this study are included in the taxon IDs.

Code availability. The code used in this analysis is included as a Supplementary Information file. The authors place no restriction on its use.

Data Collection. BLASTP was performed against the NCBI NR Database using a series of SQMO, OSC, SHC, and UbiH/COQ6 protein queries with an e-value cutoff of 10e-5 (Accession IDs: WP_027157849.1, NP_033296.1, AAH51055.1, WP_027156910.1, NP_666118.1, WP_027157848.1, WP_027157848.1). The resulting hits were vetted for conserved domains using PFAM [25]. Because of the high overlap between OSC and SHC BLAST hits, vetted results from the two searches were combined, sequences were aligned with ClustalO [26], and a tree was made with RaxML [27], using an LG matrix and 100 rapid bootstraps. The results from this tree-building exercise clearly demarcated OSC homologs from SHC, and we annotated the data accordingly.

Once the NCBI genes were vetted and annotated, subsampling was performed using custom scripts. The script appends taxonomic information to each sequence from NCBI, based on GI number, and then tabulates presence/absence data for all genes across all taxa. Taxa are divided into clades based on a chosen taxonomic rank (in our scenario, Order); if taxonomic data is missing from the NCBI taxonomy, the script automatically looks one rank deeper. The script then determines whether any taxon in the clade contains a copy of all gene queries. If multiple taxa contain all gene queries, the program randomly selects one; otherwise, the program randomly selects a taxon with the highest number of matching homologs. If the chosen taxon is missing a gene, but other species in that clade have a copy of the gene, one sequence is randomly selected and added to the dataset. For poorly represented clades in NCBI (amoebozoans, excavates, chlorophytes,

rhodophytes, rhizarians, and alveolates) we repeated this sampling at the one-per-genus level. Still lacking rhizarian and alveolate data, we collected additional sequences from the “SAR” clade using the Marine Microbial Eukaryotic Transcriptome Sequencing Project (http://marinemicroeukaryotes.org/).

Following this process, certain taxa were removed from the dataset. Taxa were removed if they contained multiple paralogs of a gene (e.g. most higher plants; 20% of all taxa), if their higher-order taxonomy is contentious (e.g haptophytes, glomeromycotes; 4% of all taxa), or if they fell outside of their taxonomically accepted Class or Superphylum in one or more analyses (e.g. nematodes, platyhelminthes; 6% of all taxa). We recovered SHC proteins in many fungi, which suggests an interesting horizontal gene transfer event from bacteria into eukaryotes. But because of poor resolution in the bacterial portion of the SHC tree, we could not determine with confidence which lineage this transfer event occurred from, and therefore chose to exclude fungal SHC from our study. Full datasets, with the removed taxa indicated, are provided as Extended Data Figs. 1 & 2.

To analyze synteny, sequence (fasta) and annotation (gff3) files were downloaded from NCBI for the following genome BioProjects: PRJNA82779, PRJNA242456, PRJNA185587, PRJNA47603, PRJNA20997, PRJNA291650, PRJNA82927, PRJNA63053, PRJDB3104, PRJNA21, PRJEA73721, PRJNA89087, PRJNA161599, PRJNA203240, PRJNA19341, and PRJNA242472. The genomes were queried based on accession numbers from the vetted SQMO and OSC proteins. Additional gene names listed in Fig. 1B are based on the original annotation files associated with the genome.

Protein alignment and tree building. We ultimately created five curated datasets, which were aligned with ClustalO: [1] SQMO (80 taxa, 841 characters), [2] SQMO + UbiH/COQ6 (172 taxa, 1071 characters), [3] OSC (104 taxa, 1267 characters), [4] OSC + SHC (174 taxa, 1475 characters), and [5] SQMO + OSC (116 taxa, 2166 characters). Trees were built for [2] and [4] using RaxML on the CIPRES Science Gateway, with an LG substitution model and 100 rapid bootstraps. Bayesian trees were constructed using MrBayes [28] on the CIPRES Science Gateway, with MCMC sampling every 1000 generations for 4 million generations, or until convergence was reached according to the *stopval* parameter (average standard deviation of split frequencies < 0.01). RaxmL trees for datasets [2] and [4] are provided in Extended Data Figs. 3 and 4 respectively, and MrBayes trees are provided in Extended Data Figs. 5 and 6.

Molecular clocks were constructed using BEAST [29]. We chose lognormal relaxed clocks, using a yule process, an LG substitution model, and 4 gamma + invariant categories. In dataset [5] we partitioned by gene, leaving substitution and clock models unlinked, but trees linked. The clocks were calibrated using 18 fossils [14,30-32] as described in Extended Table 2. Calibration points were constrained as monophyletic, and three additional monophyly constraints were set at nodes with poor resolution in our analyses, but high resolution in previous multi-gene analyses: (1) Unikonta/Amorphea, (2) Bikonta (including Bacterial Group II) and (3) non-oomycete stramenopiles. In analyses where Excavata was treated as sister to Bikonta, we set an additional monophyly constraint joining the two clades. BEAST MCMC chains were run for 20 million generations,

sampling every 1,000 generations. We re-ran all BEAST analyses a second time to test for the reproducibility of our results, and then ran each analyses a third time sampling only from the prior (i.e. ignoring molecular data), to verify that the dates we obtained were not driven solely from fossil constraints. The results of these replications for relevant nodes are plotted in Extended Data Fig. 7A. Finally, we generated calibration-free molecular clocks using the RelTime method [33] to test the effect of fossil calibration priors on our estimated divergence times. Plots of BEAST node dates mapped against relative dates from RelTime (provided in Extended Data Fig. 7B) demonstrate a linear relationship, suggesting no single fossil calibration is significantly altering the shape of the tree.

Methods References

25 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**, D279-D285 (2016).

26 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, n/a-n/a, doi:10.1038/msb.2011.75 (2011).

27 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

28 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572-1574 (2003).

29 Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**, 1969-1973 (2012).

30 Yin, Z. *et al.* Sponge grade body fossil with cellular resolution dating 60 Myr before the Cambrian. *Proceedings of the National Academy of Sciences* **112**, E1453-E1460 (2015).

31 Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica* **18.1.1FC** (2015).

32 Gold, D. A., Runnegar, B., Gehling, J. G. & Jacobs, D. K. Ancestral state reconstruction of ontogeny supports a bilaterian affinity for Dickinsonia. *Evolution & Development* **17**, 315-324, doi:10.1111/ede.12168 (2015).

33 Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A. & Kumar, S. A Protocol for Diagnosing the Effect of Calibration Priors on Posterior Time

Estimates: A Case Study for the Cambrian Explosion of Animal Phyla. *Molecular Biology and Evolution* **32**, 1907-1912 (2015).

Extended Data Legends

Extended Data Figure 1: Maximum likelihood (RAxML) tree, showing removal of problematic SQMO sequences.

Extended Data Figure 2: Maximum likelihood (RAxML) tree, showing removal of problematic OSC sequences.

Extended Data Figure 3: Maximum likelihood (RAxML) tree from vetted SQMO dataset.

Extended Data Figure 4: Maximum likelihood (RAxML) tree from vetted OSC dataset

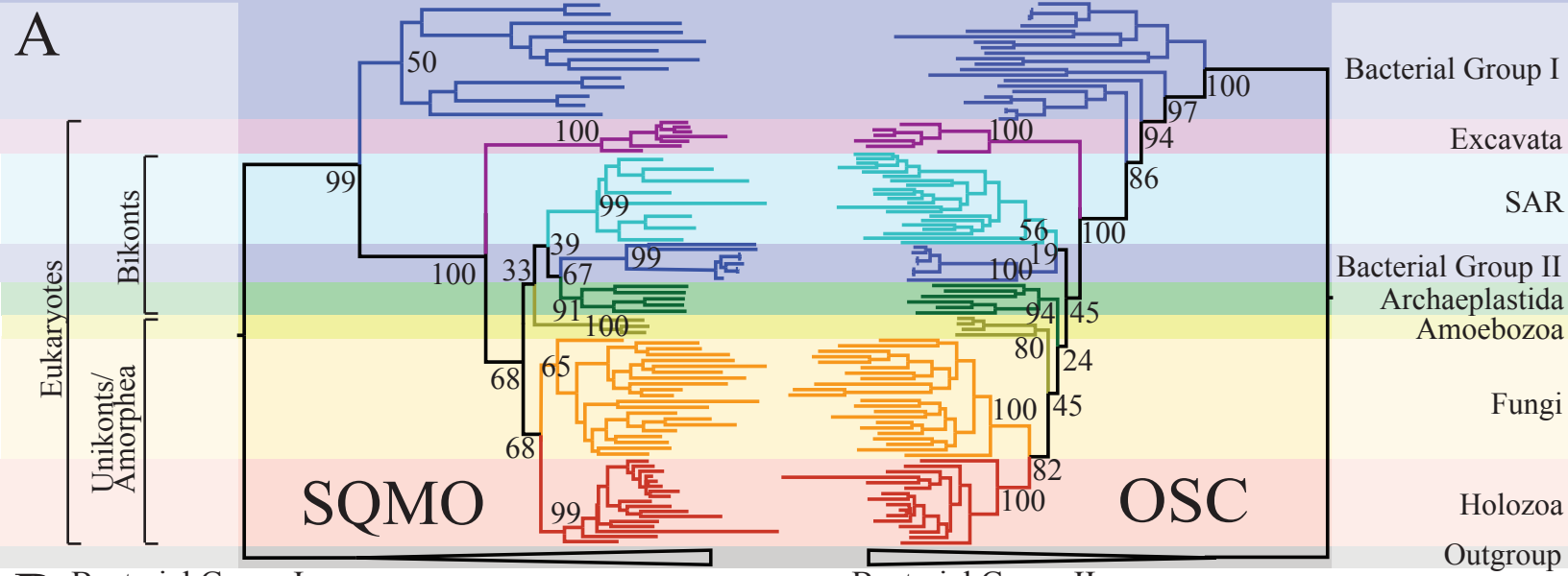
Extended Data Figure 5: Bayesian (MrBayes) tree from vetted SQMO dataset.

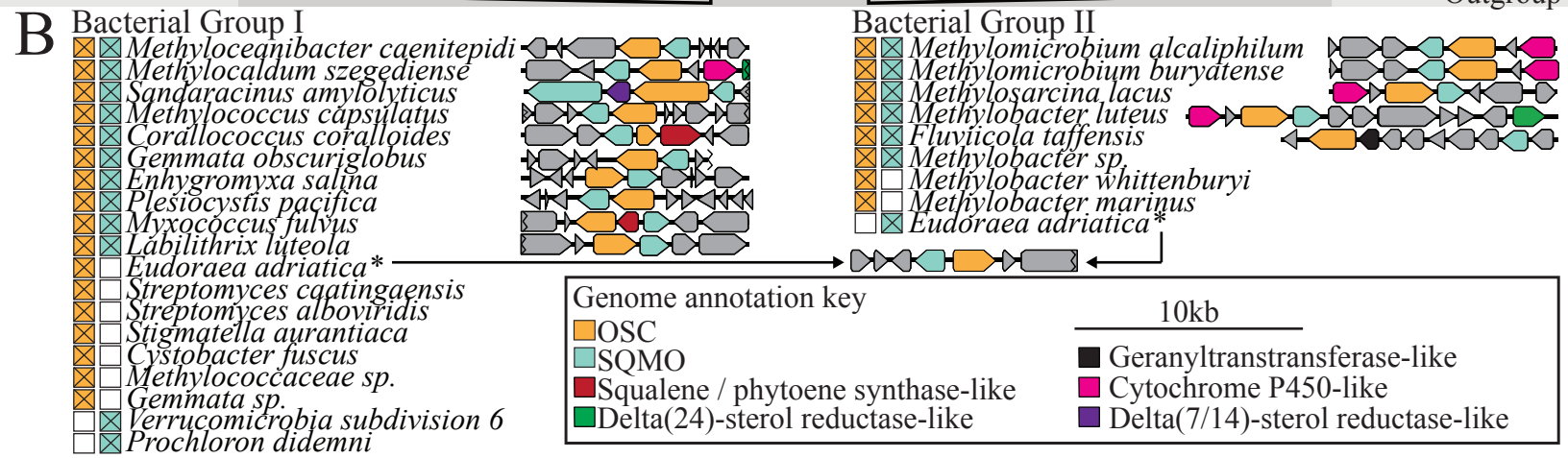
Extended Data Figure 6: Bayesian (MrBayes) tree from vetted OSC dataset.

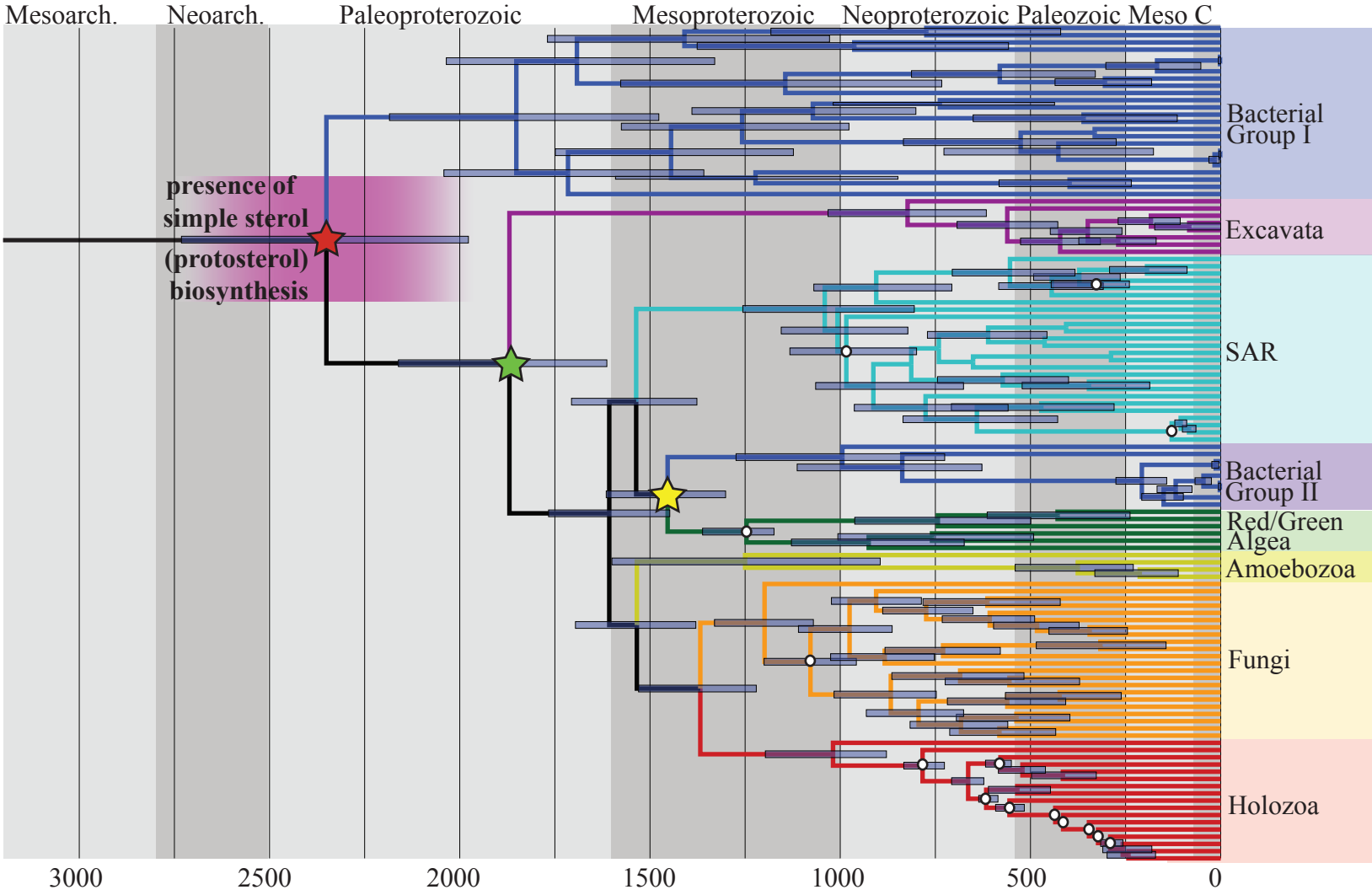
Extended Data Figure 7: Reproducibility of BEAST runs, and relationship between BEAST and RelTime trees.

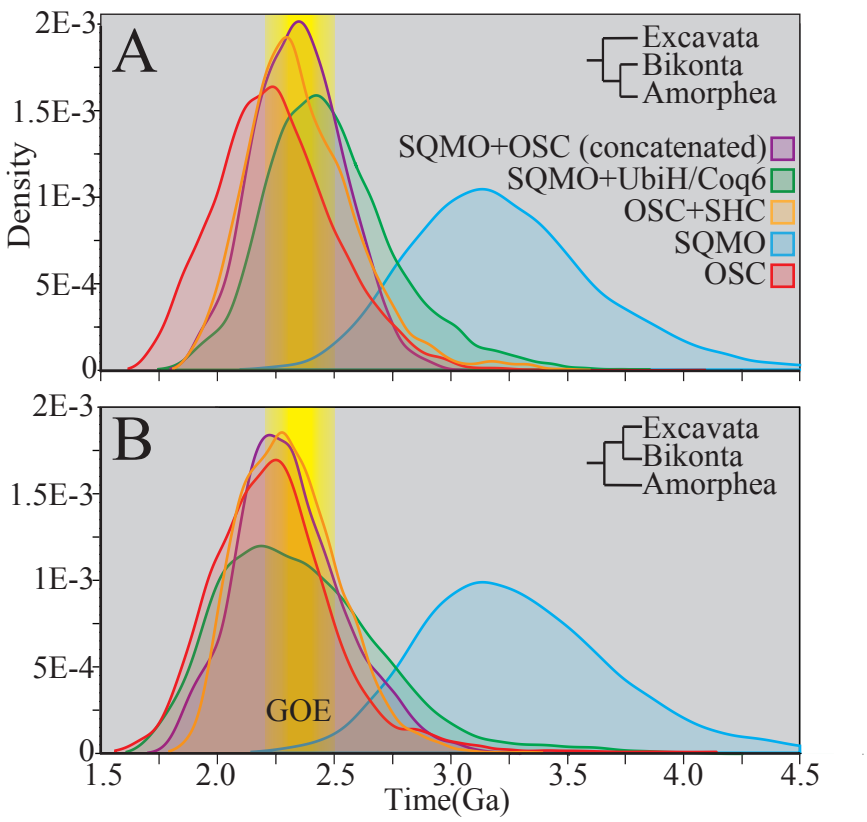
Extended Table 1: Distribution of marginal probabilities for all molecular clock analyses (as percentages), binned by geologic time.

Extended Table 2: Fossil calibration points used in molecular clock. An asterisk (*) denotes a calibration point only used in the UbiH/Coq6 outgroup. A caret (^) signifies that monophyly was not enforced on this clade as a prior in Bayesian analysis.


A
Eukaryotes
Bikonts
Unikonts/
Amorphea
SQMO
OSC
Bacterial Group I
Excavata
SAR
Bacterial Group II
Archaeplastida
Amoebozoa
Fungi
Holozoa
Outgroup
50
100
99
100
99
39
99
100
33
67
91
100
68
65
68
99
100
97
94
86
100
56
19
100
94
45
80
24
100
45
82
100


B
Bacterial Group I
Methyloceanibacter caenitepidi
Methylocaldum szegediense
Sandaracinus amylolyticus
Methylococcus capsulatus
Corallococcus coralloides
Gemmata obscuriglobus
Enhygromyxa salina
Plesiocystis pacifica
Myxococcus fulvus
Labilithrix luteola
Eudoraea adriatica*
Streptomyces caatingaensis
Streptomyces alboviridis
Stigmatella aurantiaca
Cystobacter fuscus
Methylococcaceae sp.
Gemmata sp.
Verrucomicrobia subdivision 6
Prochloron didemni
Bacterial Group II
Methylomicrobium alcaliphilum
Methylomicrobium buryatense
Methylosarcina lacus
Methylobacter luteus
Fluviicola taffensis
Methylobacter sp.
Methylobacter whittenburyi
Methylobacter marinus
Eudoraea adriatica*
Genome annotation key
OSC
SQMO
Squalene / phytoene synthase-like
Delta(24)-sterol reductase-like
10kb
Geranyltranstransferase-like
Cytochrome P450-like
Delta(7/14)-sterol reductase-like


Mesoarch.
Neoarch.
Paleoproterozoic
Mesoproterozoic
Neoproterozoic
Paleozoic
Meso
C
presence of simple sterol (protosterol) biosynthesis
Bacterial Group I
Excavata
SAR
Bacterial Group II
Red/Green Algea
Amoebozoa
Fungi
Holozoa
3000
2500
2000
1500
1000
500
0


A
B
Excavata
Bikonta
Amorphea
SQMO+OSC (concatenated)
SQMO+UbiH/Coq6
OSC+SHC
SQMO
OSC
Density
2E-3
1.5E-3
1E-3
5E-4
0
GOE
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Time(Ga)