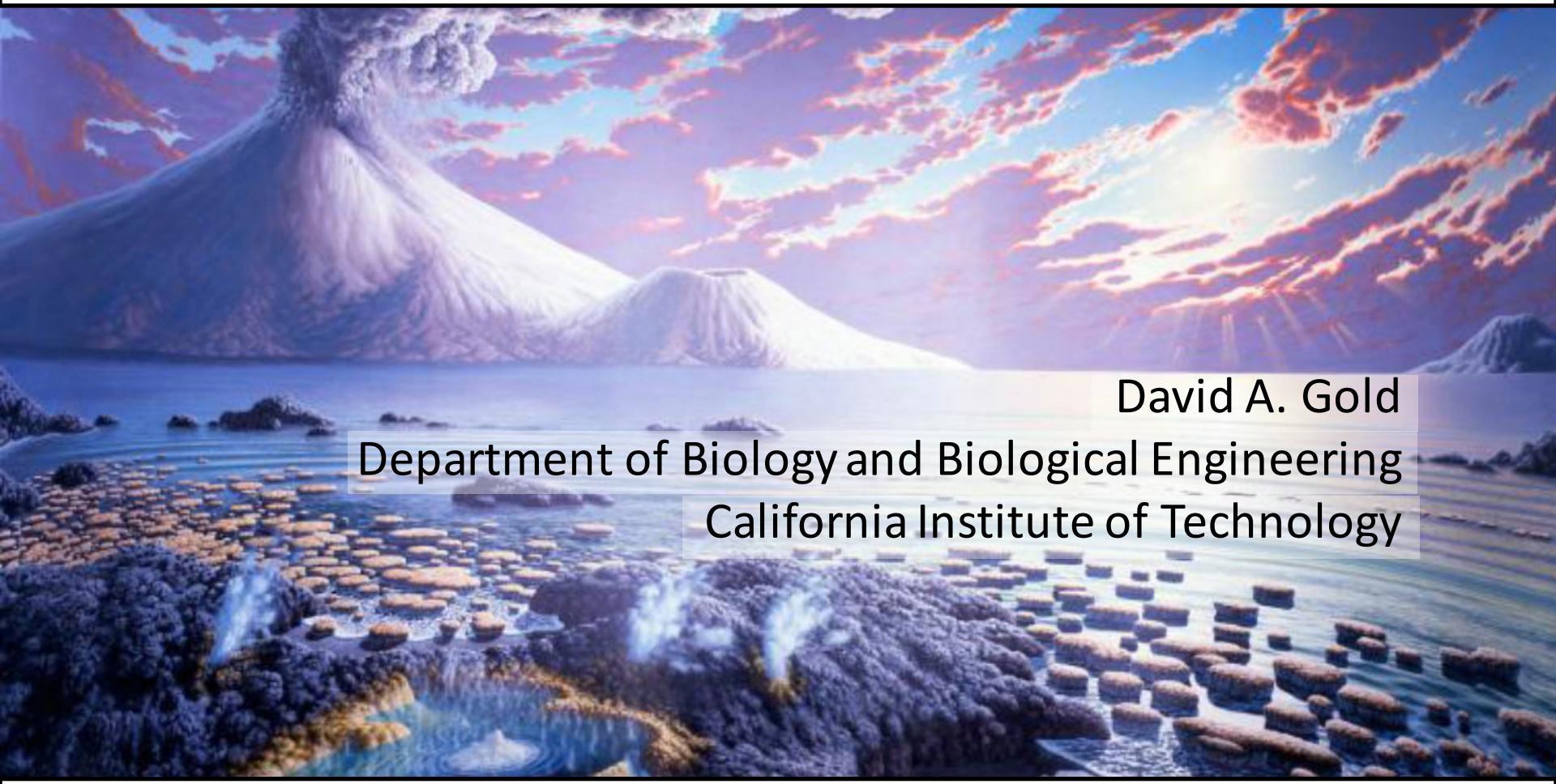
Sterol Biosynthesis and the Rise of Oxygen



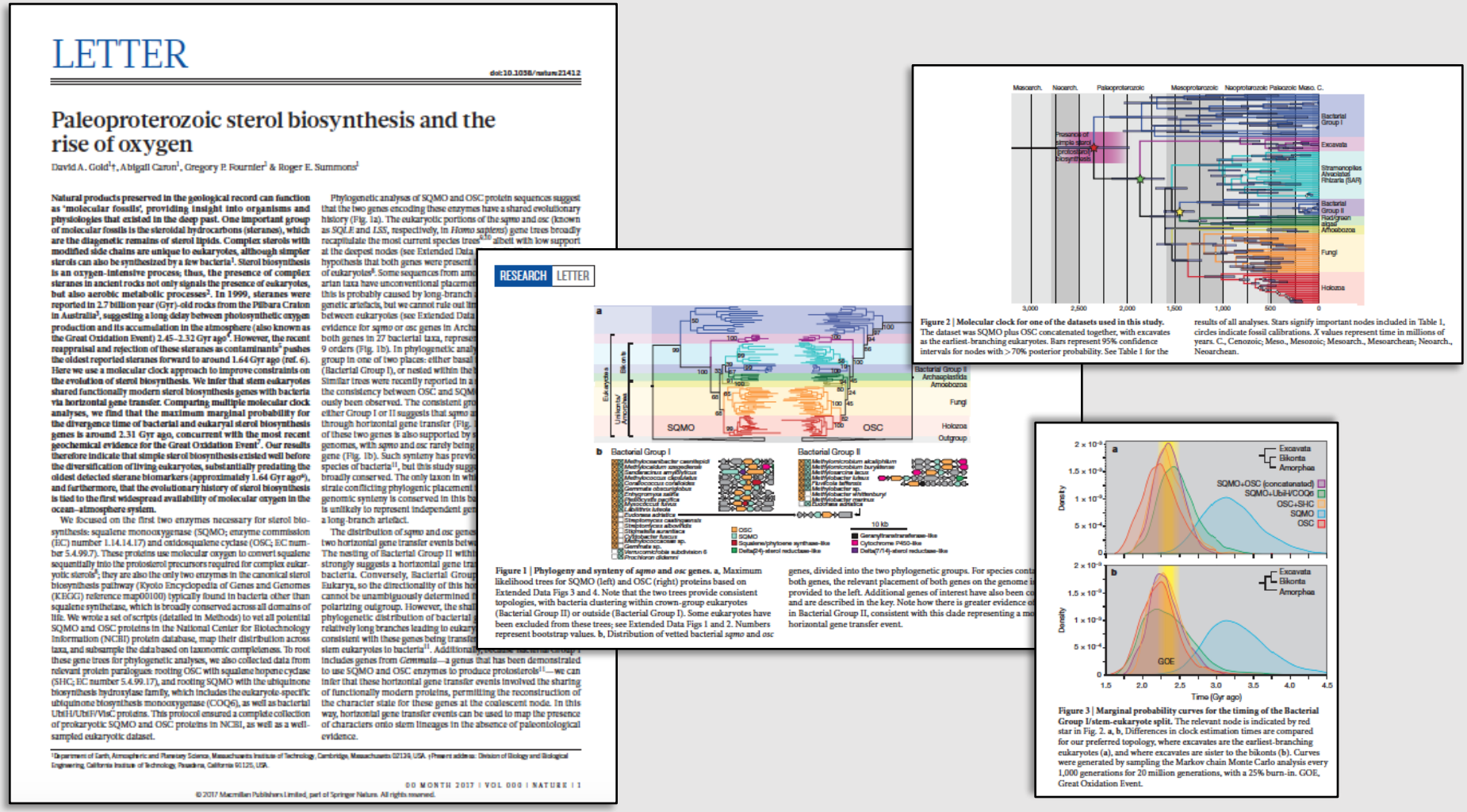
David A. Gold
Department of Biology and Biological Engineering
California Institute of Technology

The National Academies of SCIENCES ENGINEERING MEDICINE

MIT Caltech

Goal

Walk you through our recent publication in *Nature*



LETTER

doi:10.1038/nature21412

Paleoproterozoic sterol biosynthesis and the rise of oxygen

David A. Gold[1]†, Abigail Caron[1], Gregory P. Fournier[2] & Roger E. Summons[1]

Natural products preserved in the geological 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 steroid 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 (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 around 1.64 Gyr ago (ref. 6). Here 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 around 2.31 Gyr ago, concurrent with the most recent geochemical evidence for the Great Oxidation Event[4]. Our results therefore indicate that simple sterol biosynthesis existed well before the diversification of living eukaryotes, substantially predating the oldest detected sterane biomarkers (approximately 1.64 Gyr ago[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.

We focused on the first two enzymes necessary for sterol biosynthesis: squalene monooxygenase (SQMO; enzyme commission (EC) number 1.14.14.17) and oxidosqualene cyclase (OSC; EC number 5.4.99.7). These proteins use molecular oxygen to convert squalene sequentially into the protosterol precursors required for complex eukaryotic sterols[8]; they are also the only two enzymes in the canonical sterol biosynthesis pathway (Kyoto Encyclopedia of Genes and Genomes (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 Methods) to vet all potential SQMO and OSC proteins in the National Center for Biotechnology Information (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 paralogues: rooting OSC with squalene hopene cyclase (SHC; EC number 5.4.99.17), and rooting SQMO with the ubiquinone biosynthesis hydroxylase family, which includes the eukaryote-specific ubiquinone biosynthesis monooxygenase (COQ6), as well as bacterial UbiH/UbiF/VisC proteins. This protocol ensured a complete collection of prokaryotic SQMO and OSC proteins in NCBI, as well as a well-sampled eukaryotic dataset.

Phylogenetic analysis of SQMO and OSC protein sequences suggest that the two genes encoding these enzymes have a shared evolutionary history (Fig. 1a). The eukaryotic portions of the *sqmo* and *osc* (known as *SQLE* and *LSS*, respectively, in *Homo sapiens*) gene trees broadly recapitulate the most current species trees[9,10] albeit with low support at the deepest nodes (see Extended Data ... hypothesis that both genes were present ... of eukaryotes[8]. Some sequences from ... arian taxa have unconventional placeme... this is probably caused by long-branch ... genetic artefacts, but we cannot rule out lin... between eukaryotes (see Extended Data ... evidence for *sqmo* or *osc* genes in Archa... both genes in 27 bacterial taxa, represe... 9 orders (Fig. 1b). In phylogenetic analy... group in one of two places: either basal ... (Bacterial Group I), or nested within the ... Similar trees were recently reported in a ... the consistency between OSC and SQM... ously been observed. The consistent gro... either Group I or II suggests that *sqmo* a... through horizontal gene transfer (Fig. ... of these two genes is also supported by ... genomes, with *sqmo* and *osc* rarely being ... gene (Fig. 1b). Such synteny has previ... species of bacteria[11], but this study sugg... broadly conserved. The only taxon in wh... strate conflicting phylogenetic placement ... genomic synteny is conserved in this ba... is unlikely to represent independent ge... a long-branch artefact.

The distribution of *sqmo* and *osc* genes ... two horizontal gene transfer events betw... The nesting of Bacterial Group II within ... strongly suggests a horizontal gene tra... bacteria. Conversely, Bacterial Group ... Eukarya, so the directionality of this ho... cannot be unambiguously determined f... polarizing outgroup. However, the shal... phylogenetic distribution of bacterial ... relatively long branches leading to eukary... consistent with these genes being transfe... stem eukaryotes to bacteria[11]. Additional ... includes genes from *Gemmatimonadetes*—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.

[1]Department of Earth, Atmospheric and Planetary Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. †Present address: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California 91125, USA.

00 MONTH 2017 | VOL 000 | NATURE | 1

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

RESEARCH LETTER

Figure 1 | Phylogeny and synteny of *sqmo* and *osc* genes. a, Maximum likelihood trees for SQMO (left) and OSC (right) proteins 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. Numbers represent bootstrap values. **b,** Distribution of vetted bacterial *sqmo* and *osc* genes, divided into the two phylogenetic groups. For species conta... both genes, the relevant placement of both genes on the genome i... provided to the left. Additional genes of interest have also been c... and are described in the key. Note how there is greater evidence o... in Bacterial Group II, consistent with this clade representing a m... horizontal gene transfer event.

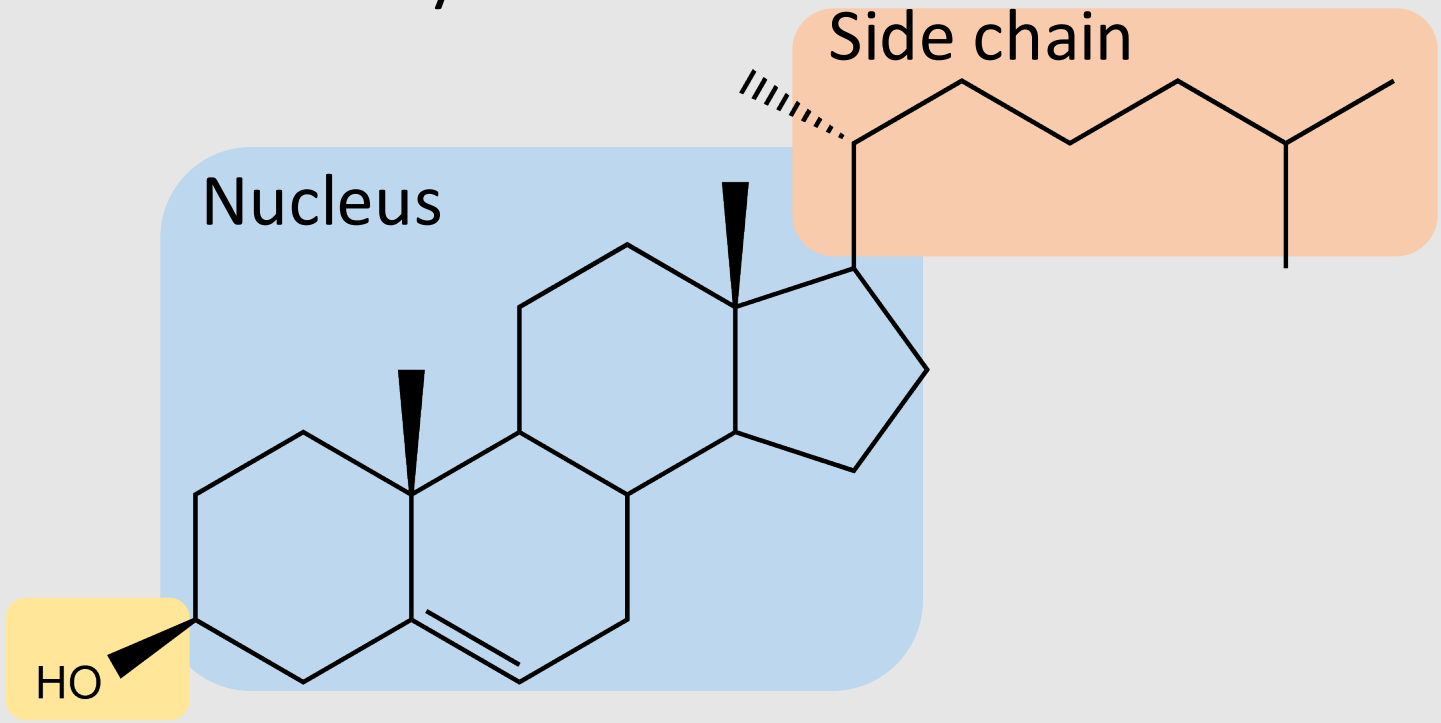
Figure 2 | Molecular clock for one of the datasets used in this study. The dataset was SQMO plus OSC concatenated together, with excavates as the earliest-branching eukaryotes. Bars represent 95% confidence intervals for nodes with >70% posterior probability. See Table 1 for the results of all analyses. Stars signify important nodes included in Table 1, circles indicate fossil calibrations. *X* values represent time in millions of years. C., Cenozoic; Meso., Mesozoic; Mesoarch., Mesoarchean; Neoarch., Neoarchean.

Figure 3 | Marginal probability curves for the timing of the Bacterial Group I/stem-eukaryote split. The relevant node is indicated by red star in Fig. 2. **a, b,** Differences in clock estimation times are compared for our preferred topology, where excavates are the earliest-branching eukaryotes (**a**), and where excavates are sister to the bikonts (**b**). Curves were generated by sampling the Markov chain Monte Carlo analysis every 1,000 generations for 20 million generations, with a 25% burn-in. GOE, Great Oxidation Event.

Conclude with the astrobiological implications

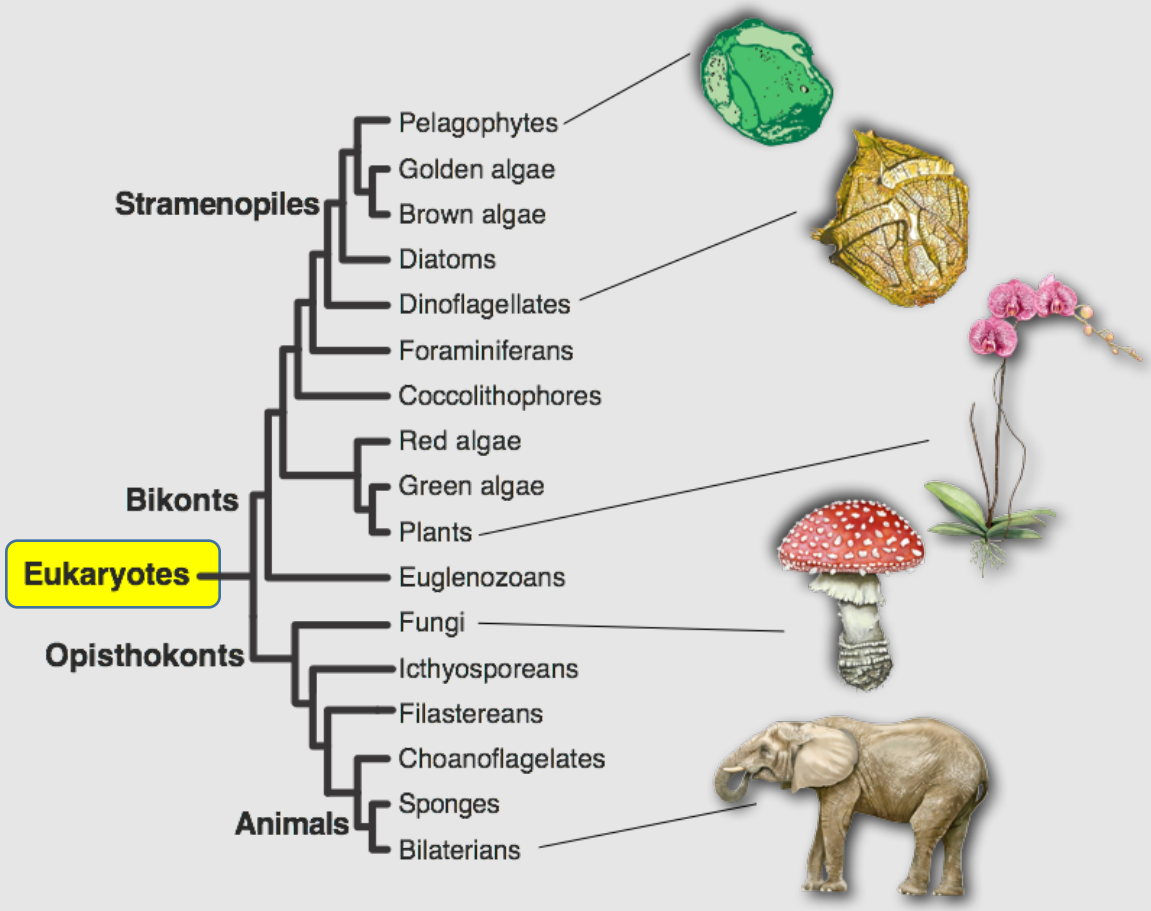
What is a sterol?

- Organic molecules and a subclass of steroids
- Play important roles in cell communication and membrane fluidity



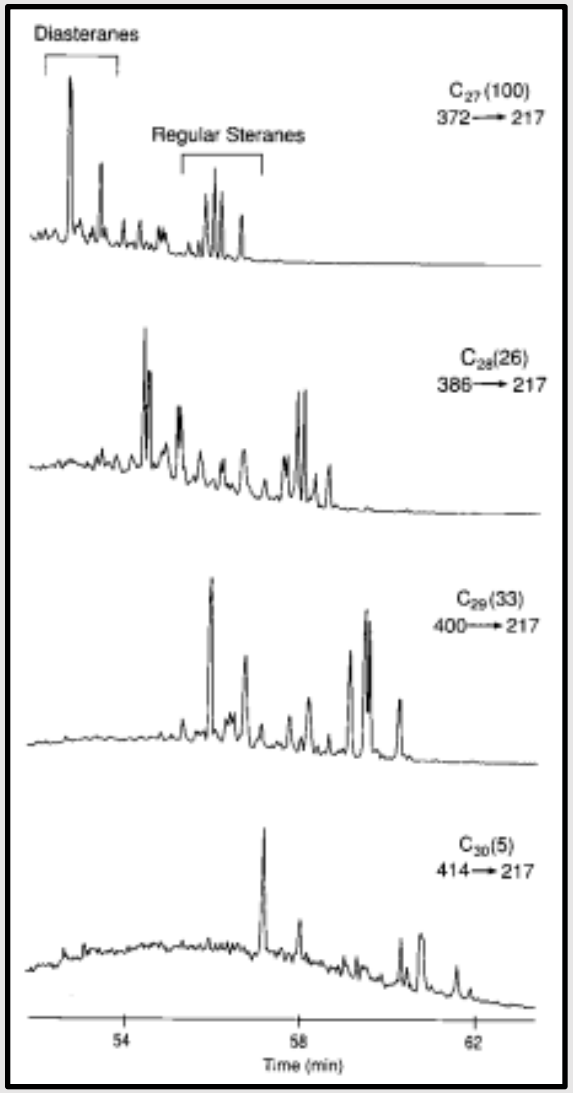
hydroxyl group

Why are sterols interesting for the origin of complex life?



- Sterols are (almost) exclusively made by eukaryotes
- Sterols require a significant amount of oxygen (11 mol O_2)
- The presence of steranes in the geological record provides evidence of oxygen and complex life

When did sterols originate?



Pilbara Craton, Australia

- Rocks from Pilbara Craton were analyzed for preserved biomolecules
- The results suggested sterols were present ~2.7 Gya

Brocks et al. S*cience* 285.5430 (1999): 1033-1036.

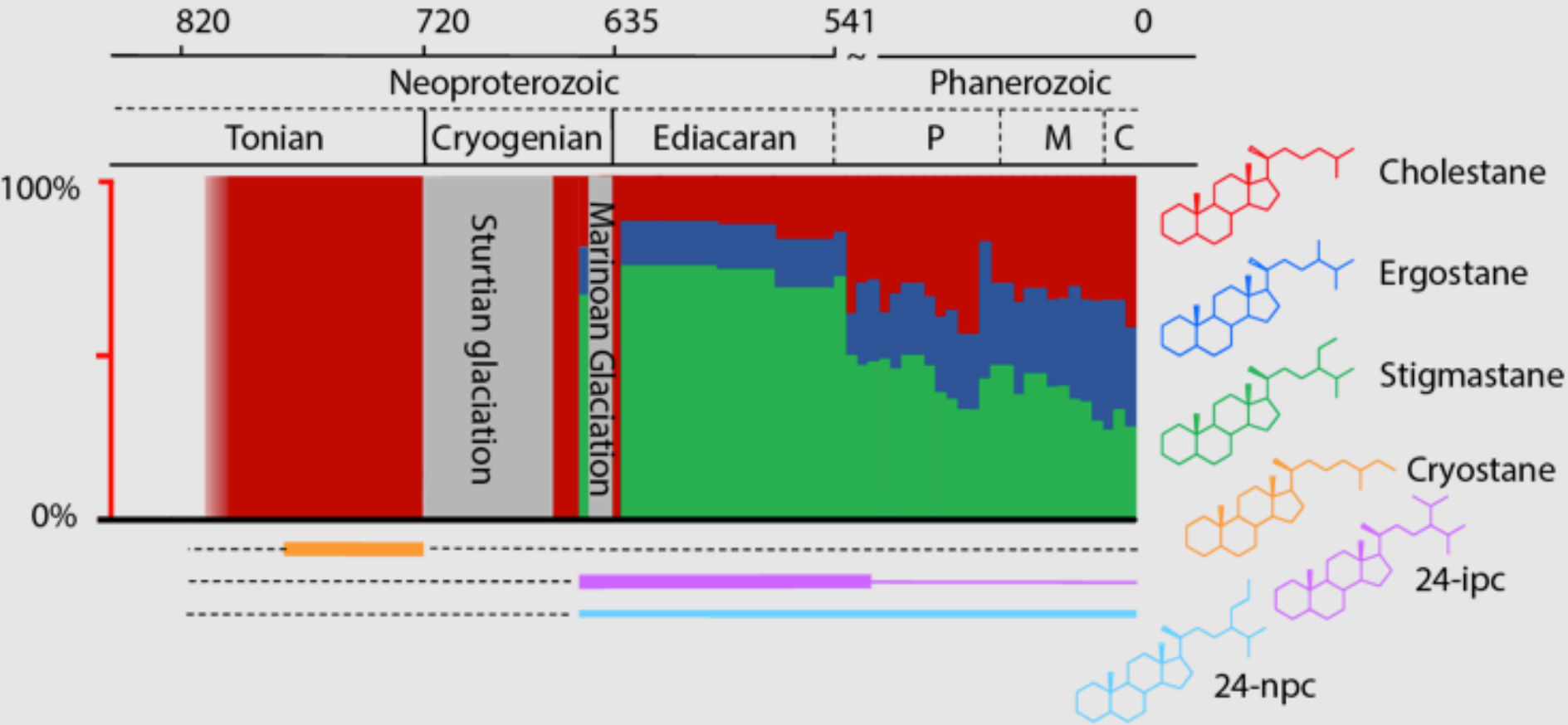
When did sterols originate?

Agouron Institute Drilling Project

- Over time, the original study has proven controversial
- The rocks were reappraised in 2015

French et al. *Proceedings of the National Academy of Sciences* 112.19 (2015): 5915-5920.

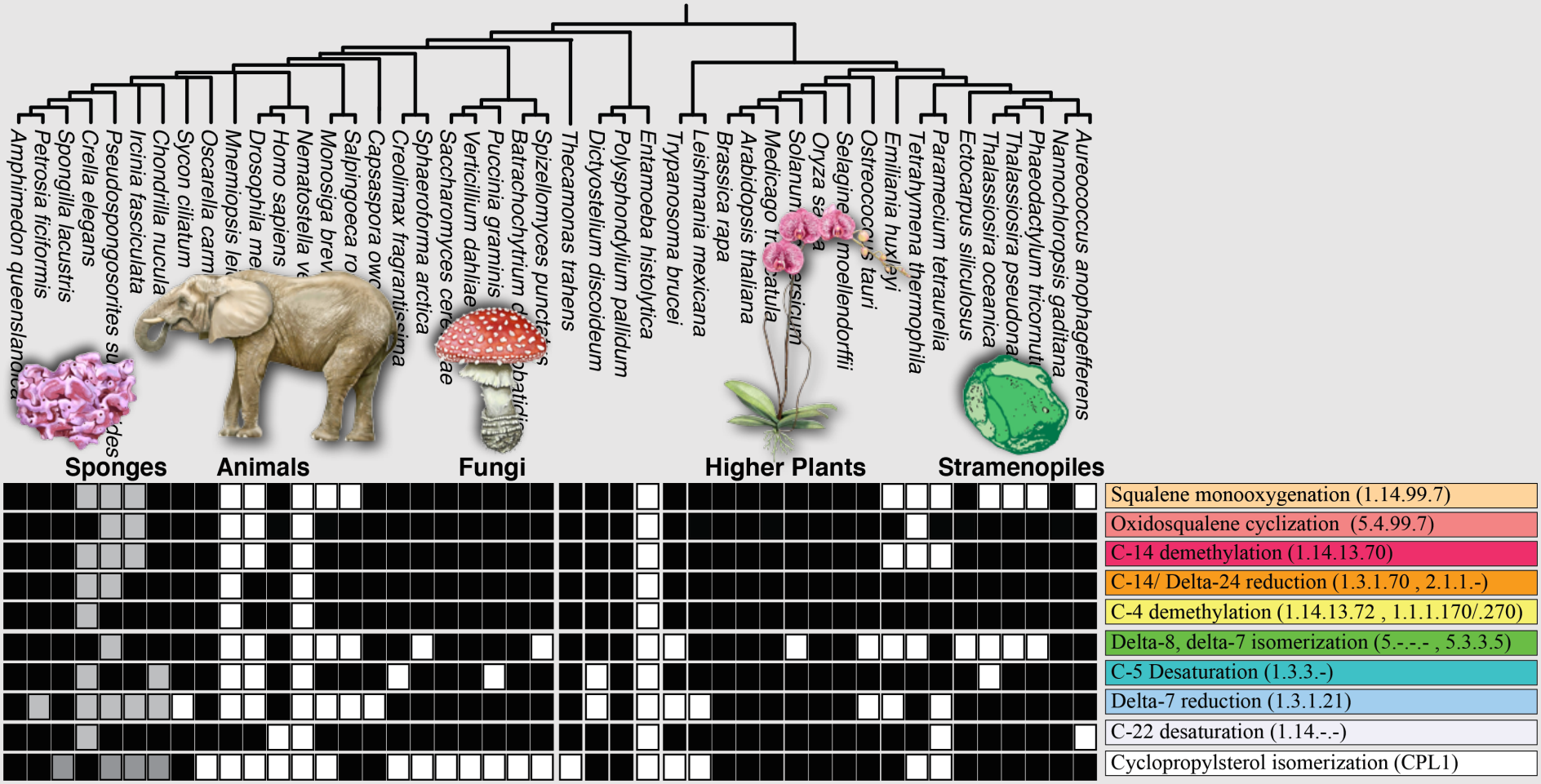
The sterane "fossil" record



Modified from Brocks et al. *Nature* (2017). *Online ahead of print*

The big question: could steranes really be that young?

Sterol production appears to be as old as the eukaryotes



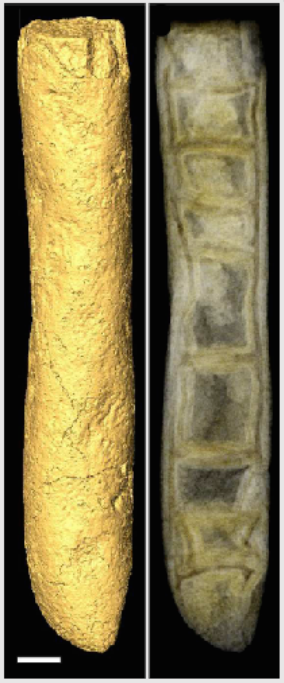
The big question: could steranes really be that young?

The fossil record suggests eukaryotes are older than 820 My

Bangiomorpha

Red algae

~1200 Mya



Rafatazmia and *Ramathallus*

Red algae?

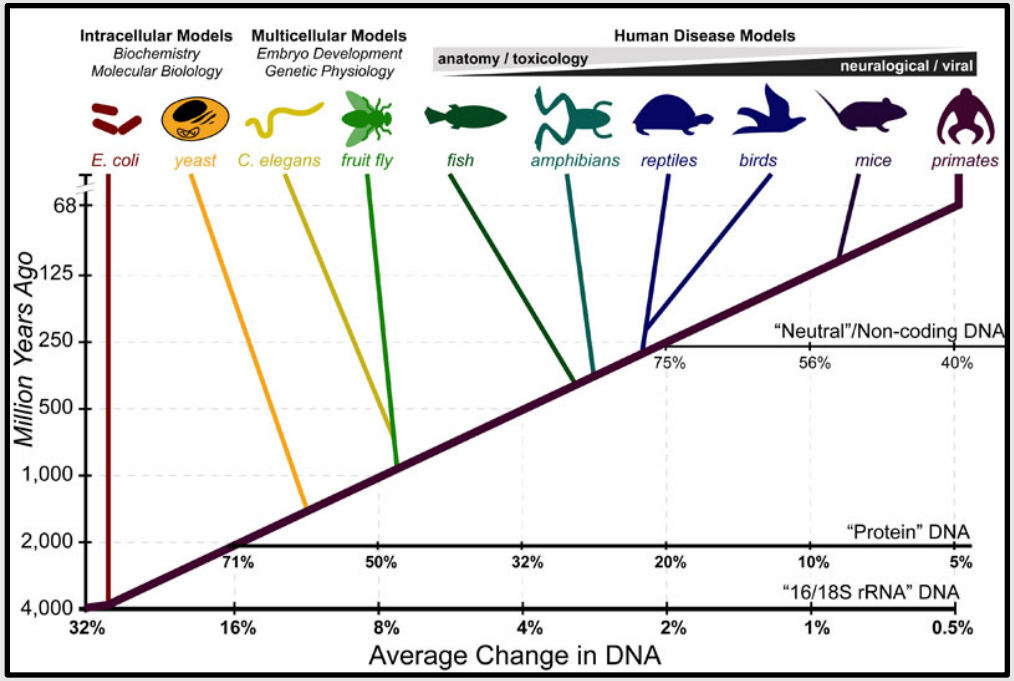
~1600 Mya

Butterfield, Nicholas J. *Paleobiology* 26.3 (2000): 386-404.
Bengtson, Stefan, et al. *PLoS biology* 15.3 (2017): e2000735.

Two ways to date the origin of life forms

Fossils

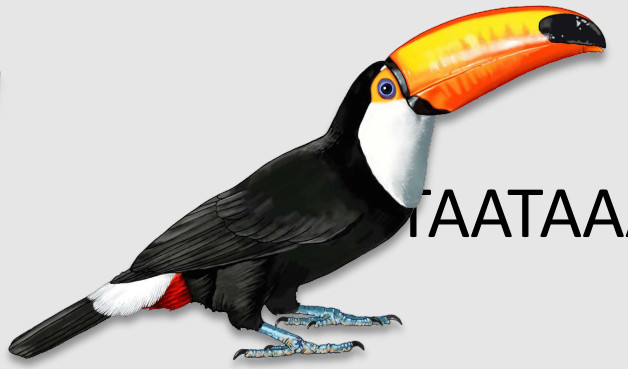
Molecular clocks



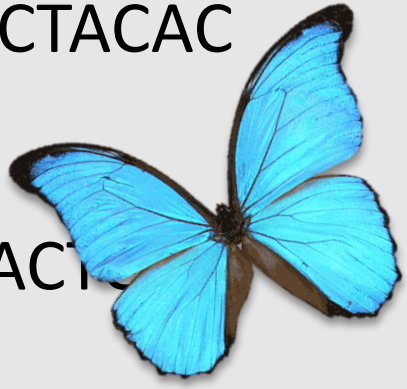
How does a molecular clock work?

Step 1: Sequence DNA

AATTAGATTACACCCC



TAATAAATAGACTACAC



CCGGAATTAGATTACT

GGGGGGCGAAATAGATTAC

TTATAAATAGACTACACTTT

How does a molecular clock work?

AATTAGATTACAC

AAATAGACTACAC

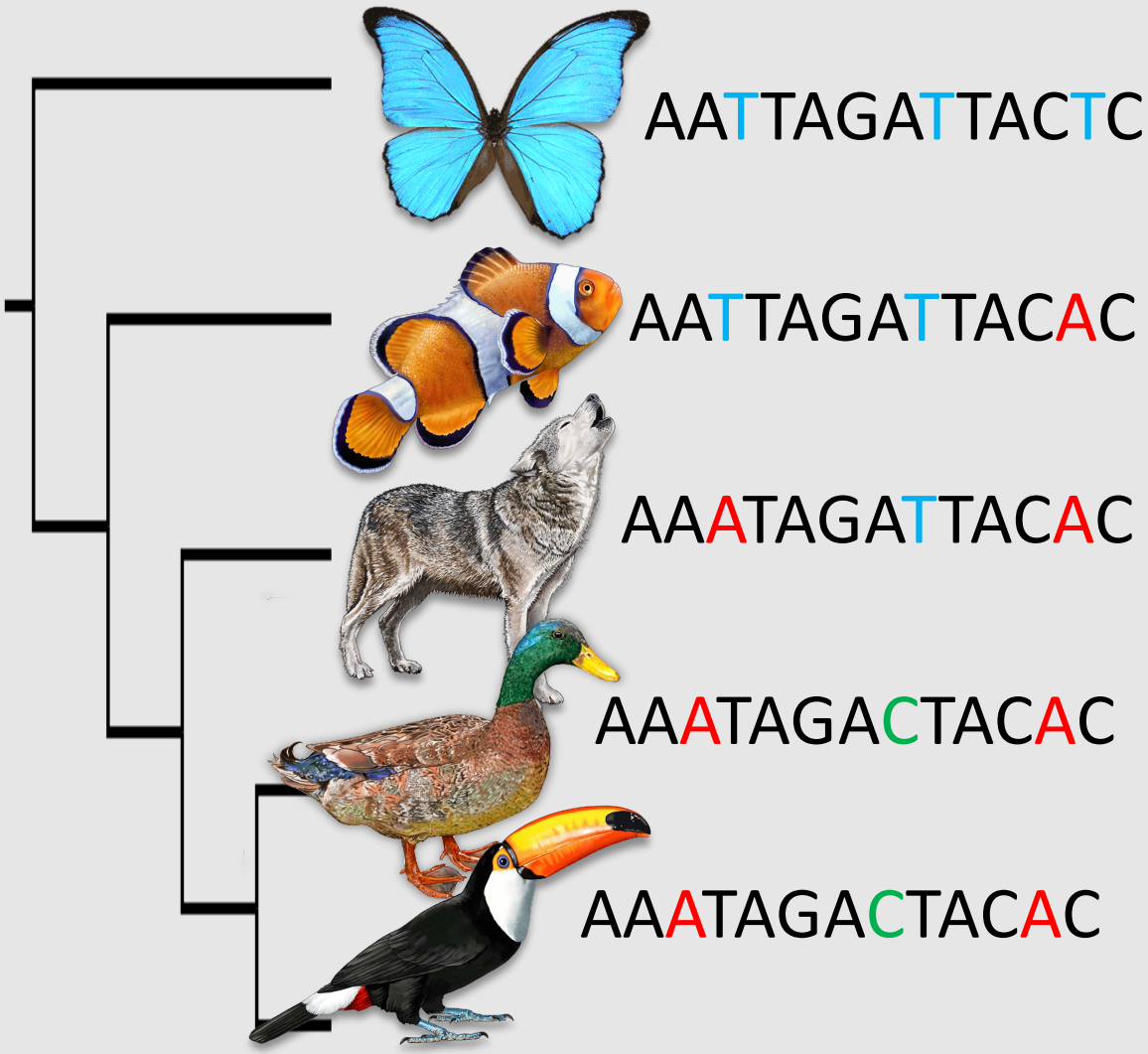
AATTAGATTACTC

AAATAGATTACAC

AAATAGACTACAC

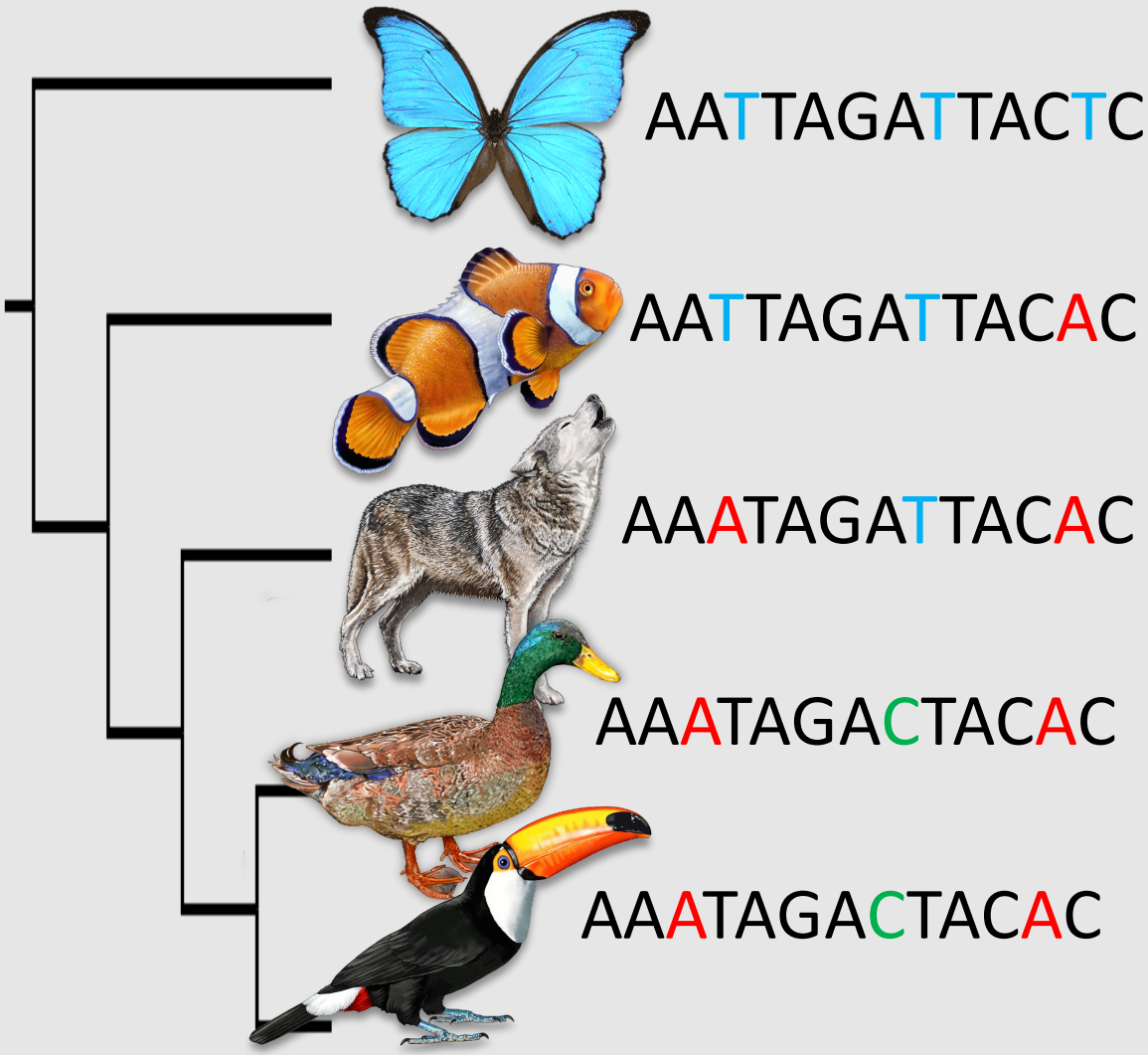
Step 2: Align regions of DNA that are homologous and look for differences (mutations)

How does a molecular clock work?



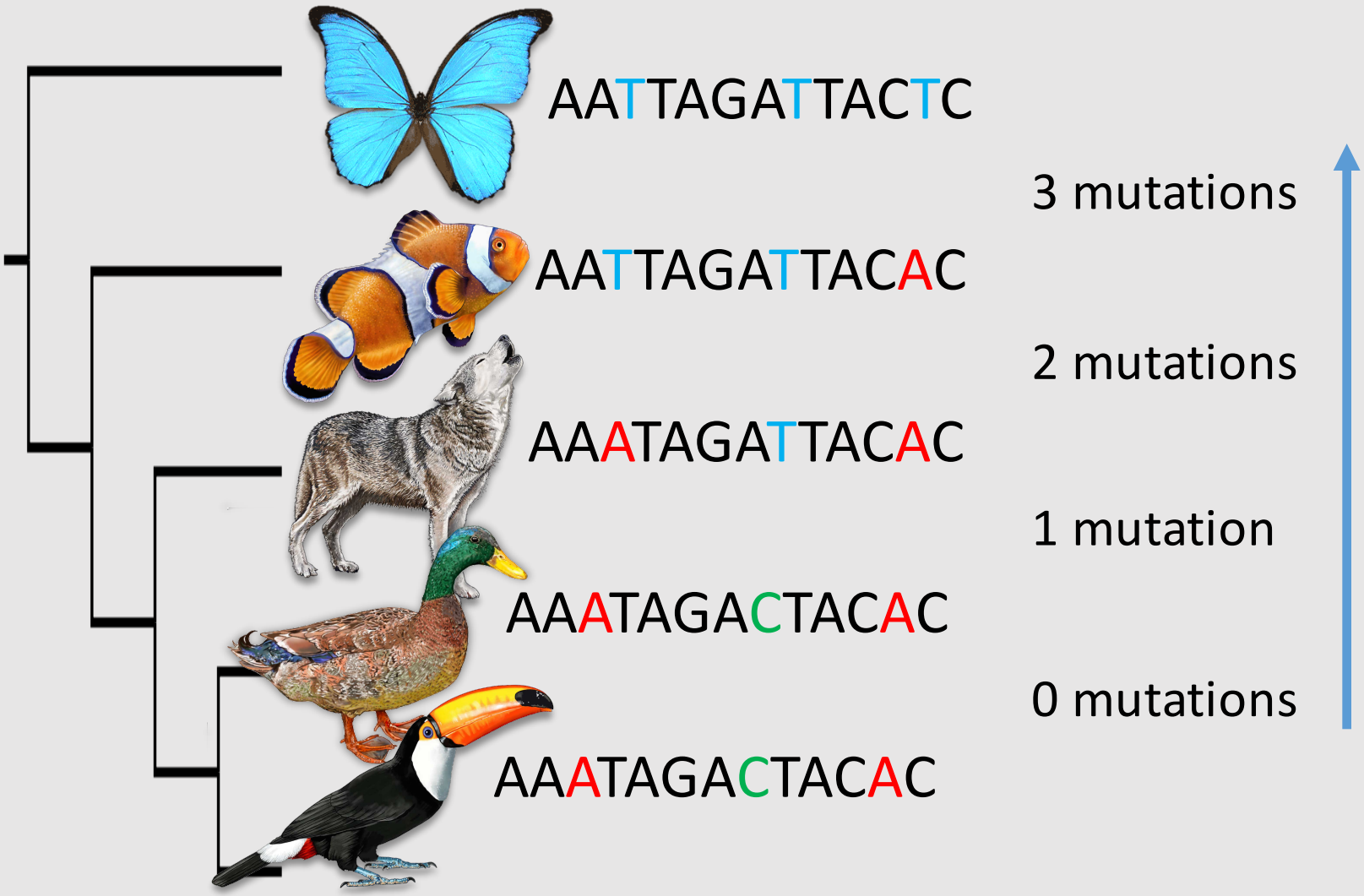
Step 3: Use this information to reconstruct an evolutionary tree (using parsimony or more complex mathematical models of DNA mutations)

How does a molecular clock work?

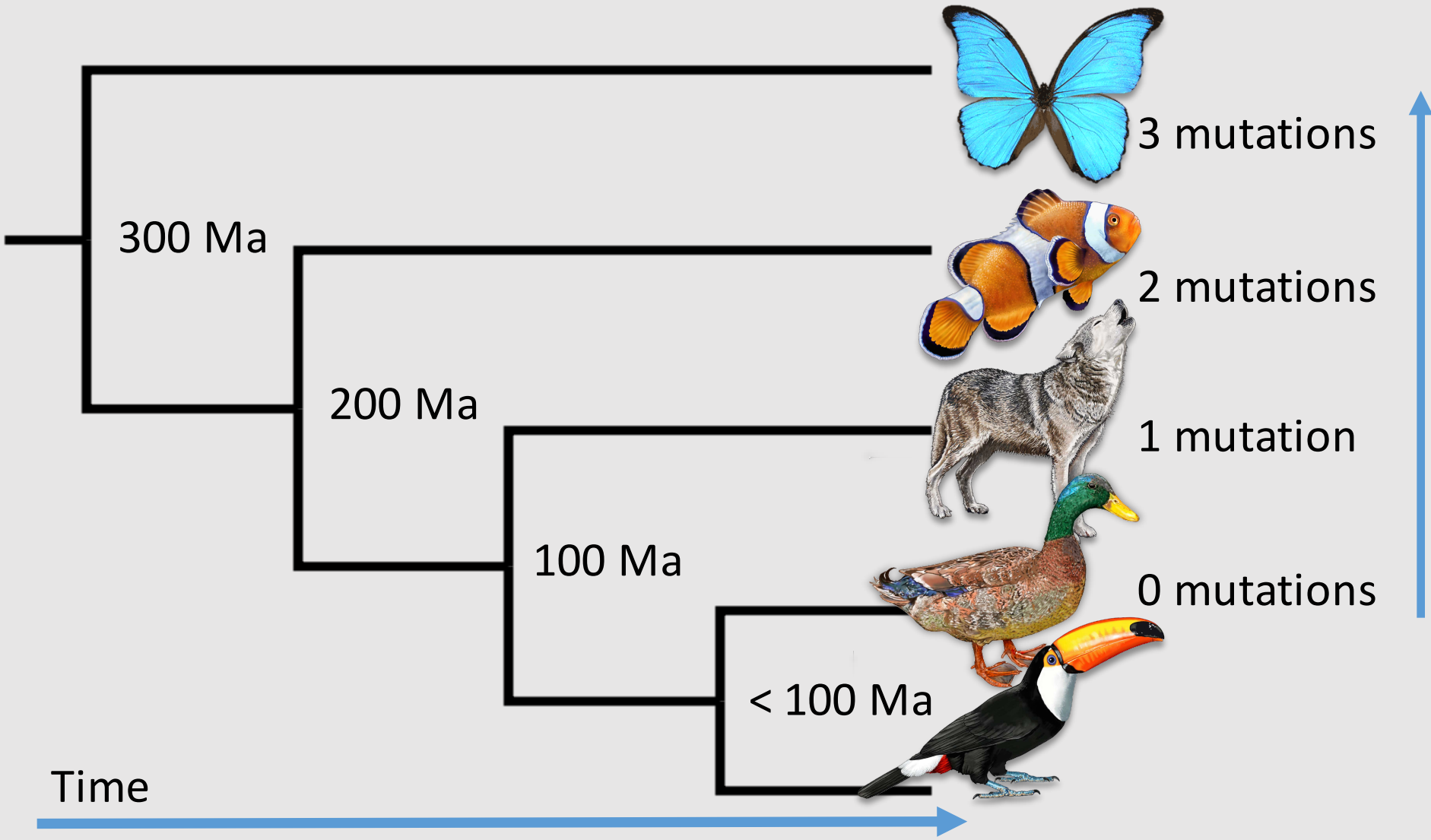


Let's assume that 1 mutation occurs every 100 million years (a strict molecular clock)

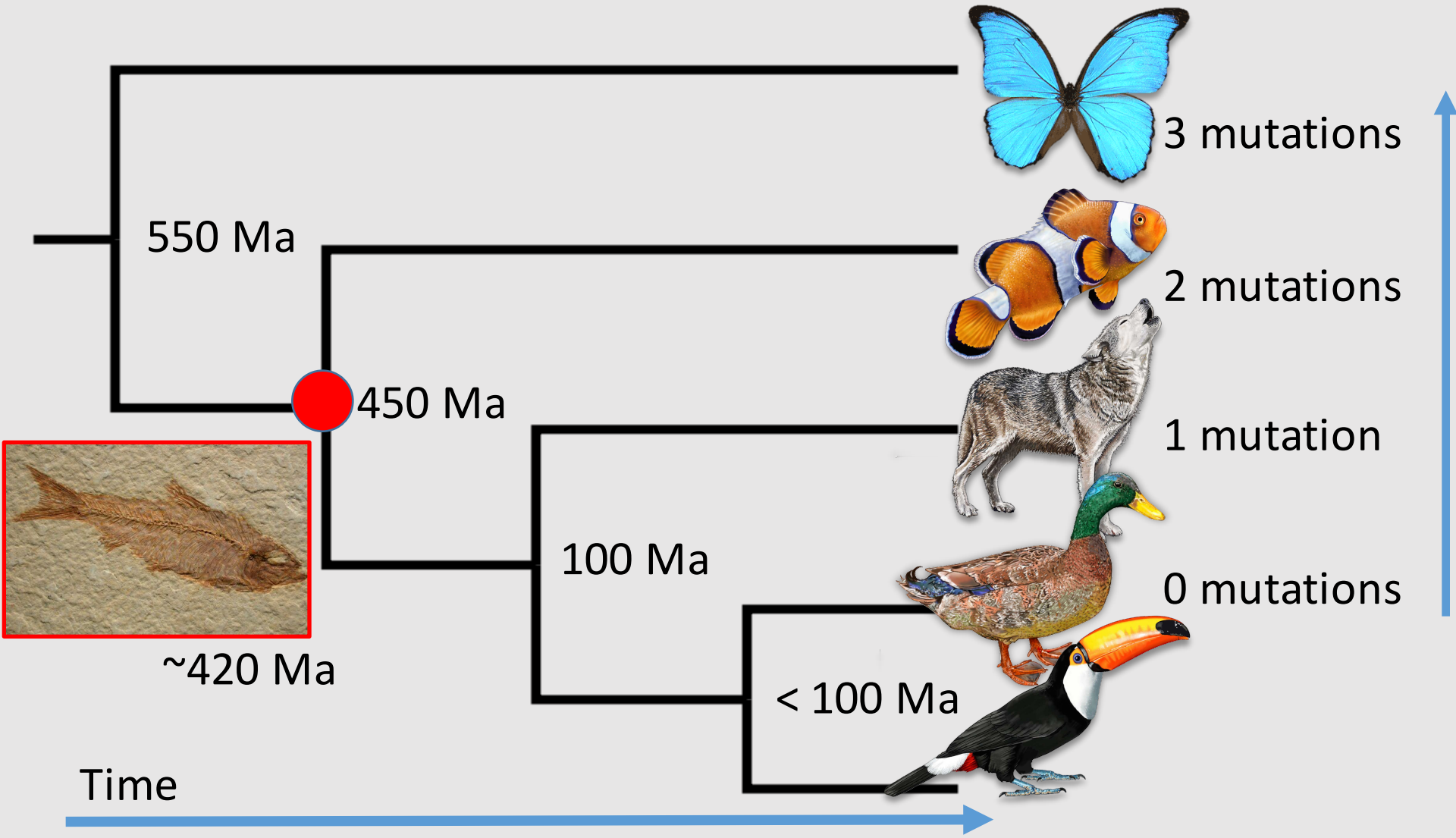
How does a molecular clock work?



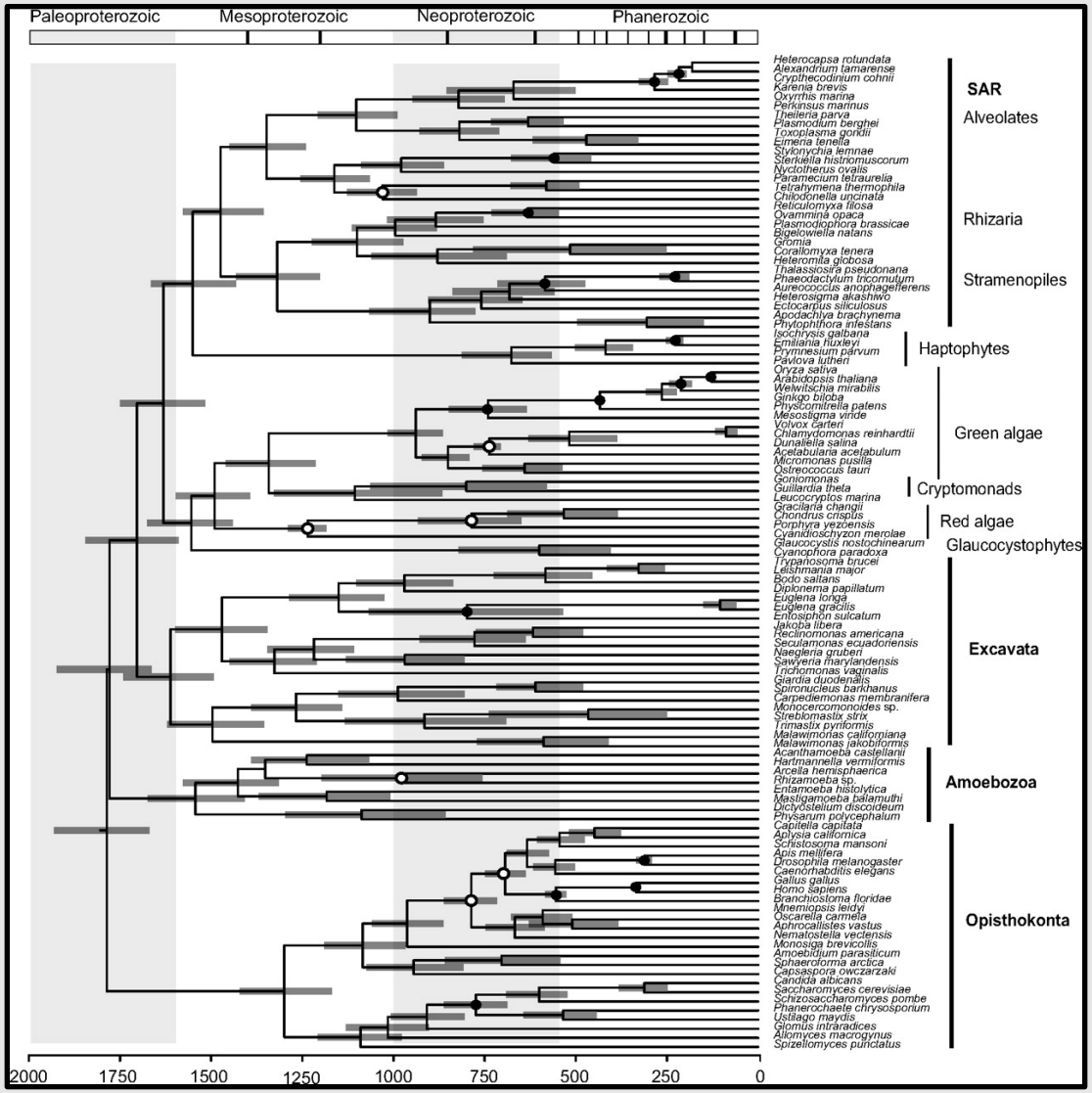
How does a molecular clock work?



How does a molecular clock work?



How does a molecular clock work?



- Molecular clock of eukaryotes
- Origin ~1.8 Gya

Parfrey et al. *Proceedings of the National Academy of Sciences* 108.33 (2011): 13624-13629.

Molecular clock dates for the origin of eukaryotes

1082.2 Mya	Gaya et al. 2015
1781.1 Mya	Parfrey et al. 2011
1624.0 Mya	Hedges et al. 2004

- These dates suggest steranes could be twice as old as their fossil record
- And they could be older…

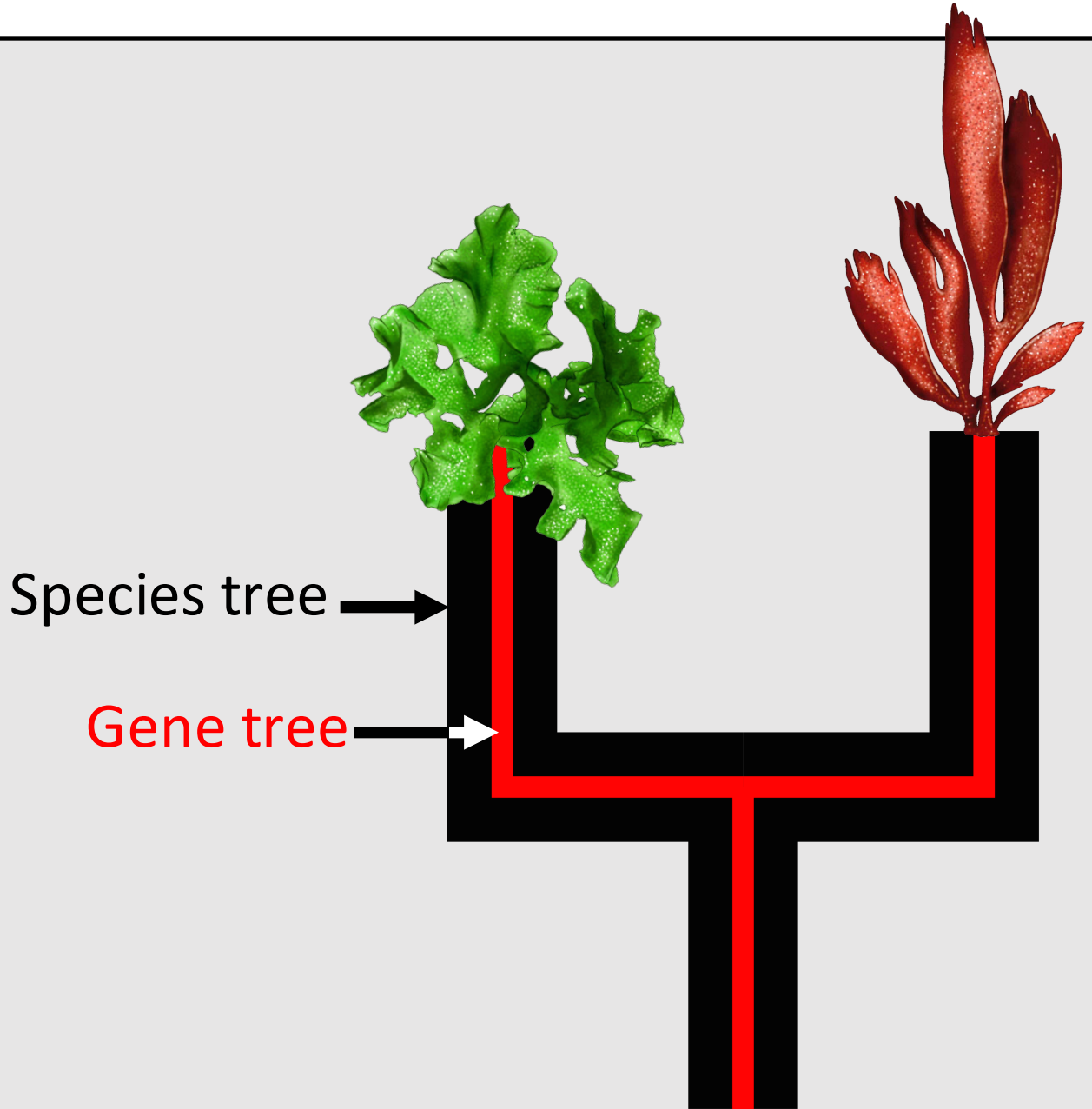
- Is there a way we can study the origin of sterols specifically?

Gaya et al. *Proceedings of the National Academy of Sciences* 112.37 (2015): 11600-11605.
Parfrey et al. *Proceedings of the National Academy of Sciences* 108.33 (2011): 13624-13629.
Hedges et al *BMC evolutionary biology* 4.1 (2004): 2.

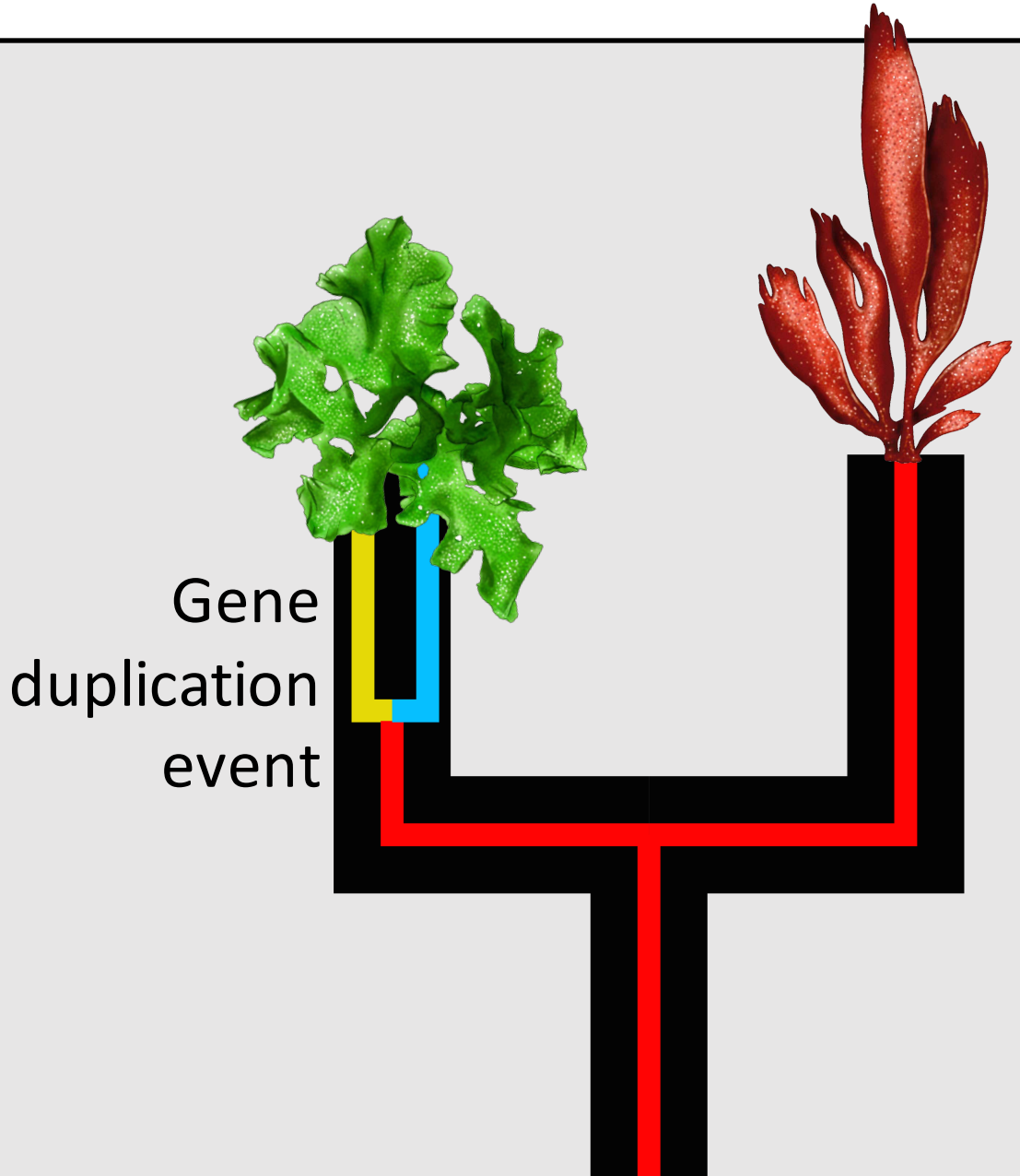
The sterol biosynthesis pathway

The evolution of genes versus species

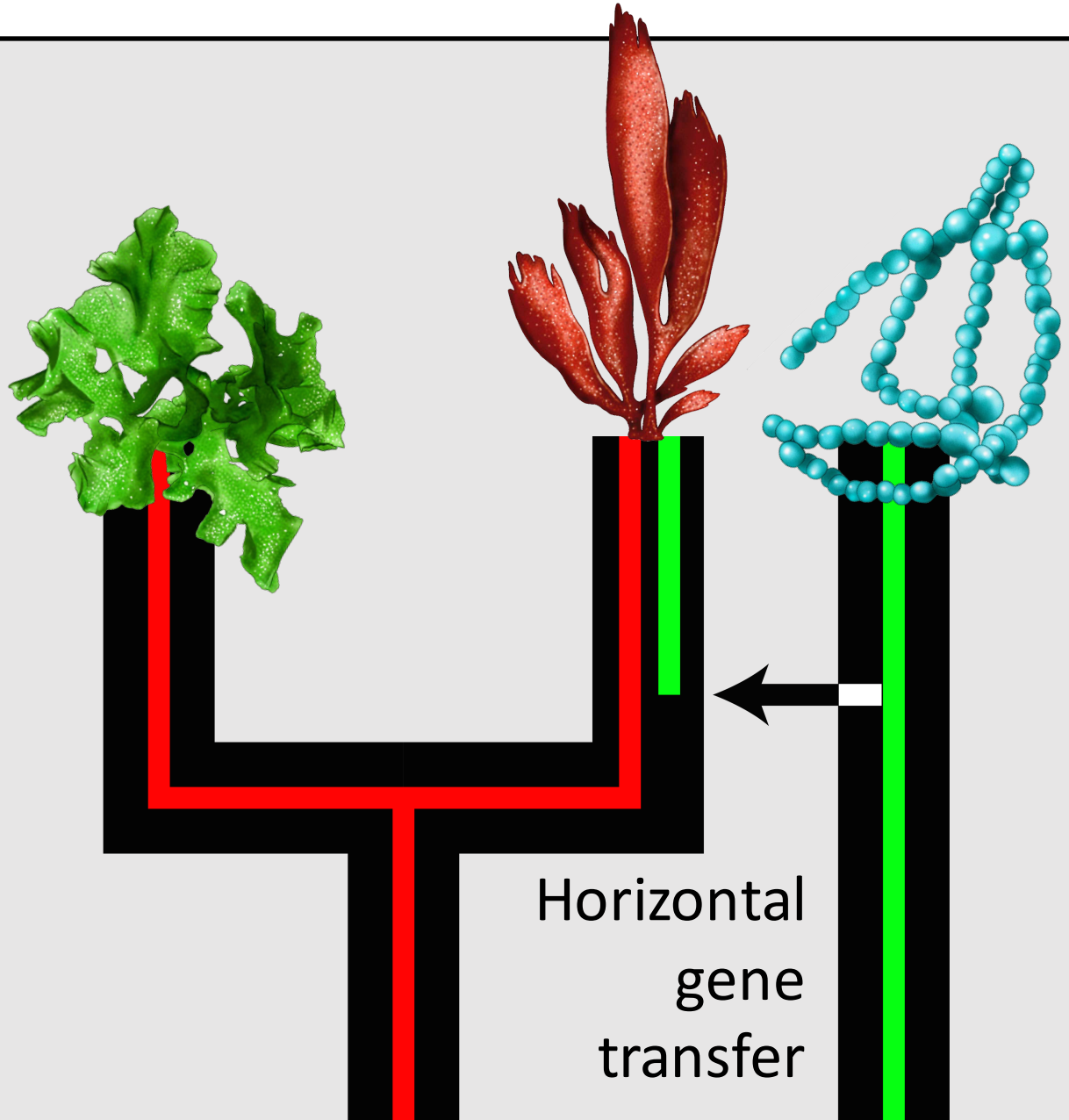
The evolution of genes versus species



The evolution of genes versus species

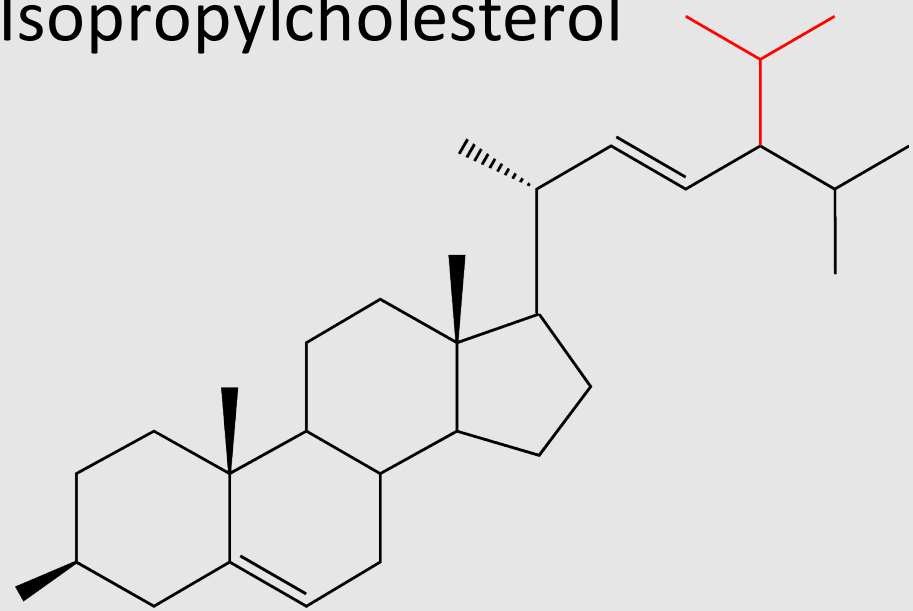


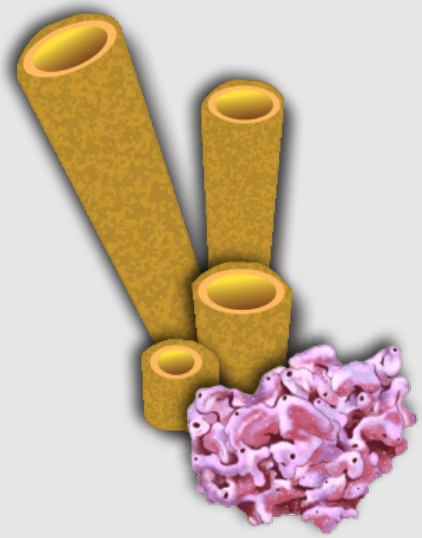
The evolution of genes versus species

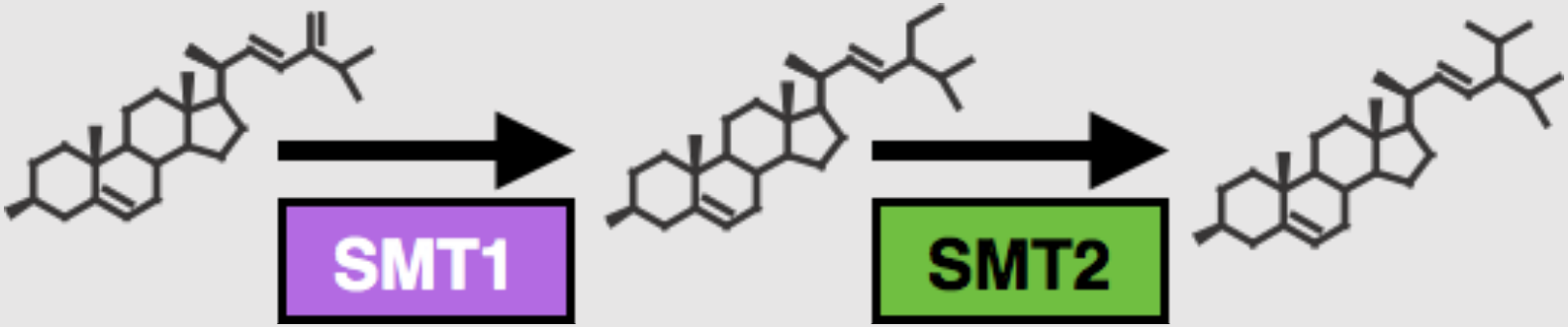


Gene-focused molecular clocks reveal the origin of 24-ipc

Isopropylcholesterol

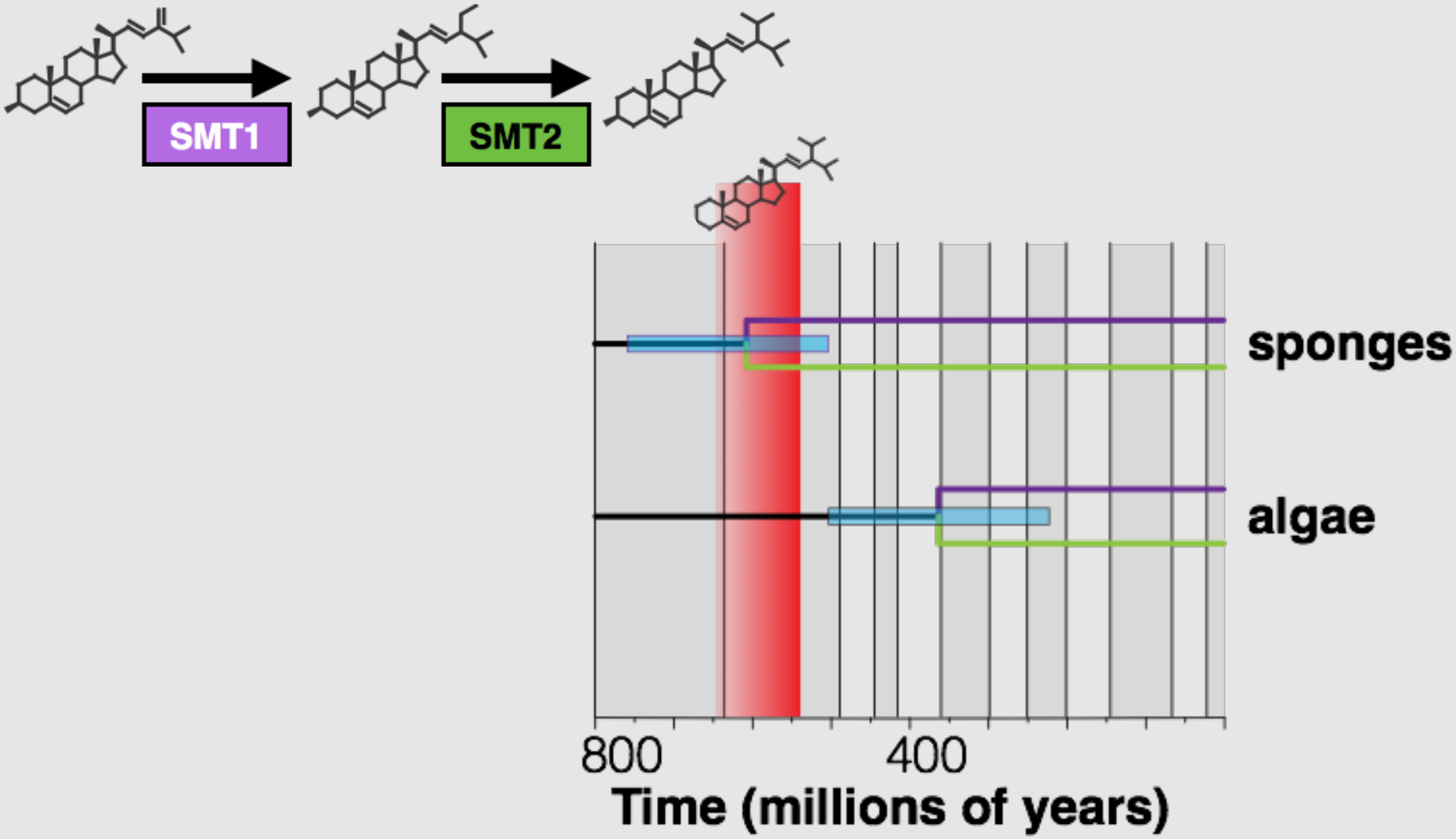






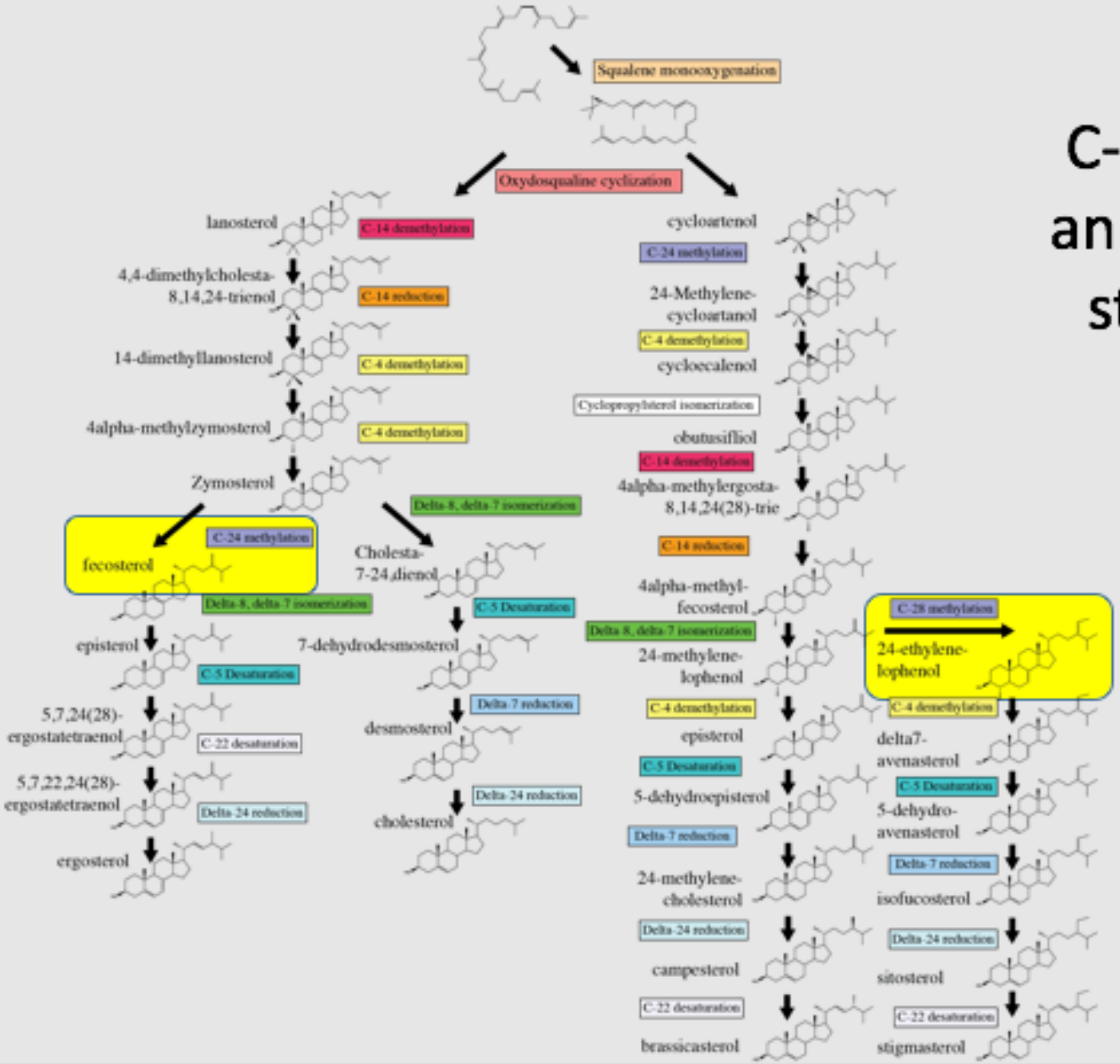
Gold, David A., et al. "Sterol and genomic analyses validate the sponge biomarker hypothesis." Proceedings of the National Academy of Sciences 113.10 (2016): 2684-2689.dos

Gene-focused molecular clocks reveal the origin of 24-ipc



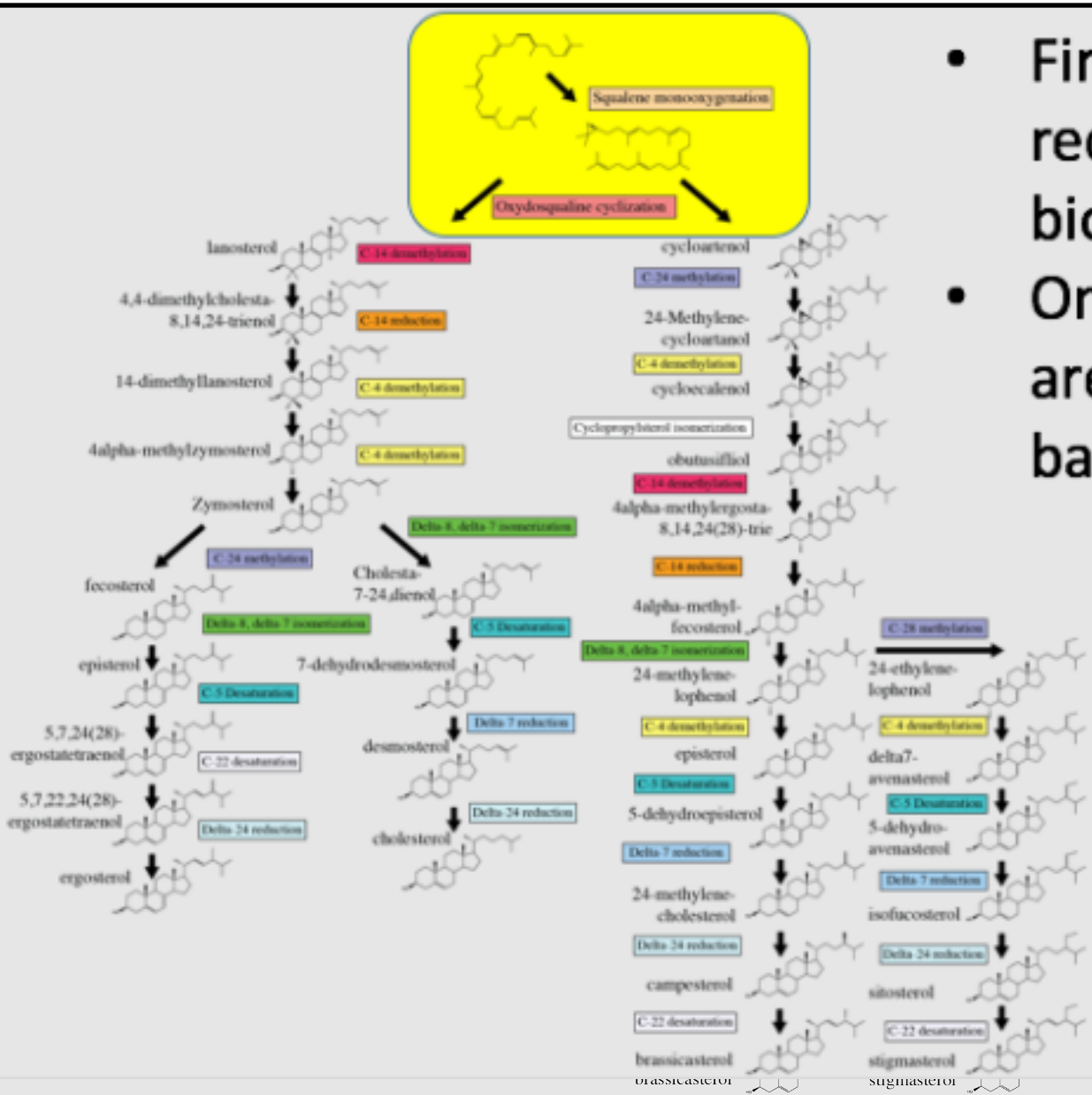
Gold, David A., et al. "Sterol and genomic analyses validate the sponge biomarker hypothesis." Proceedings of the National Academy of Sciences 113.10 (2016): 2684-2689.dos

The sterol biosynthesis pathway



C-24 methylation is an advanced step in sterol biosynthesis

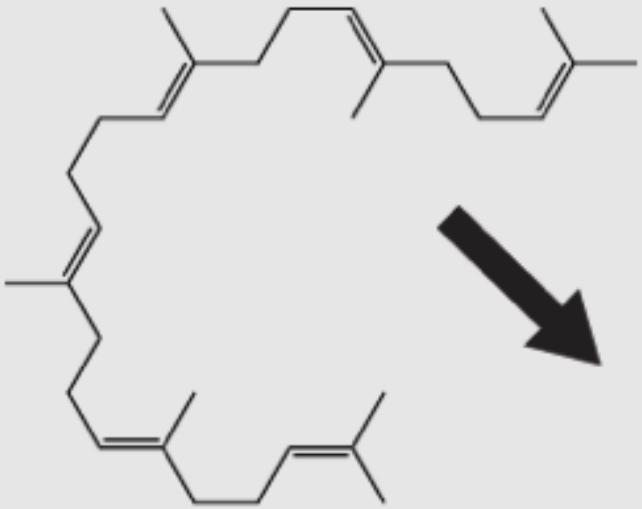
The sterol biosynthesis pathway



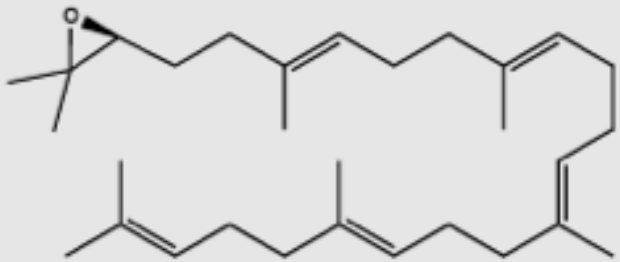
- First two genes required in all sterol biosynthesis pathways
- Only two genes that are also found in bacteria

Squalene monooxegenase (SQMO)

squalene

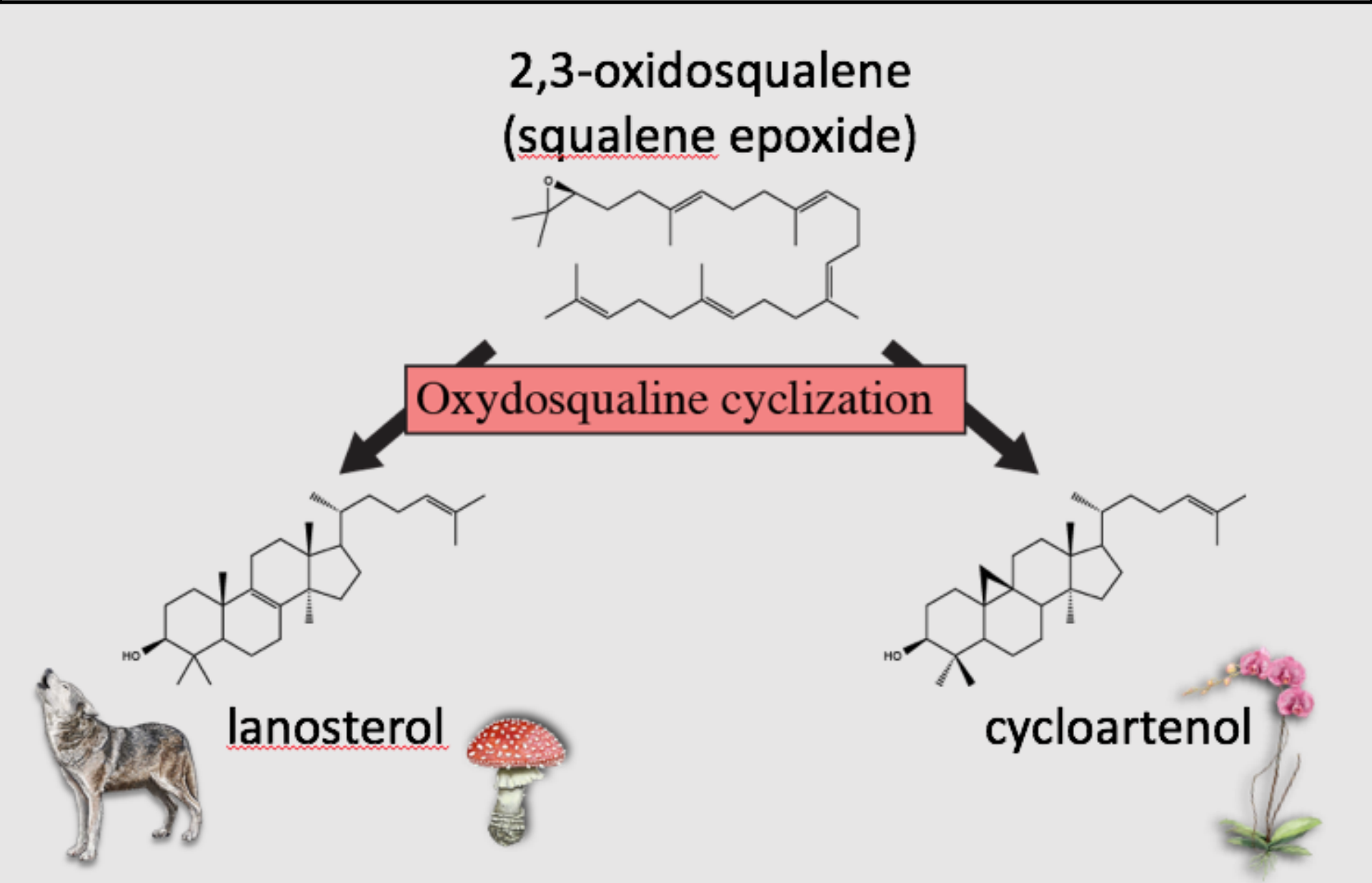


Squalene monooxygenation

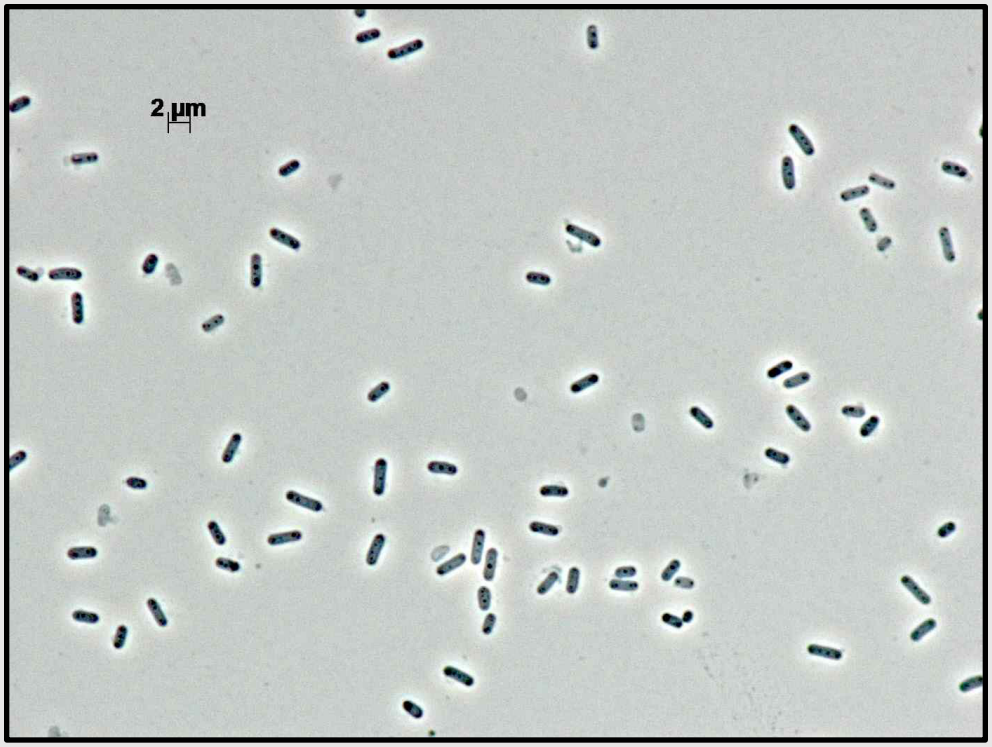


2,3-oxidosqualene
(squalene epoxide)

Oxidosqualene cyclase (OSC)



Sterol biosynthesis in bacteria



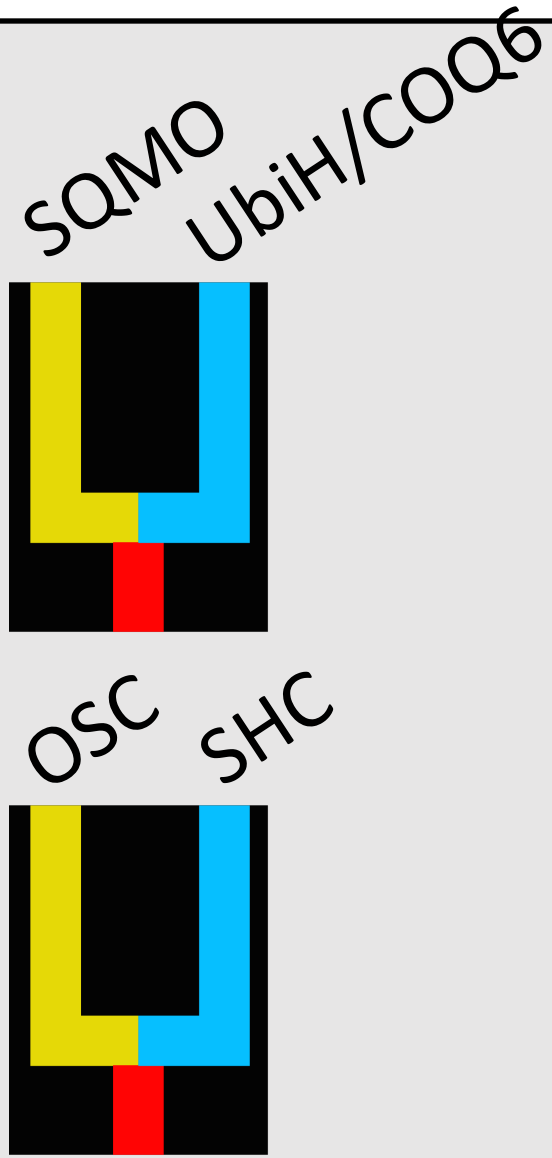
Labilithrix luteola

- Rare but found in diverse taxa
- Some taxa make cycloartenol and some make lanosterol
- Many bacteria modify their sterols, but do so with different genes than eukaryotes

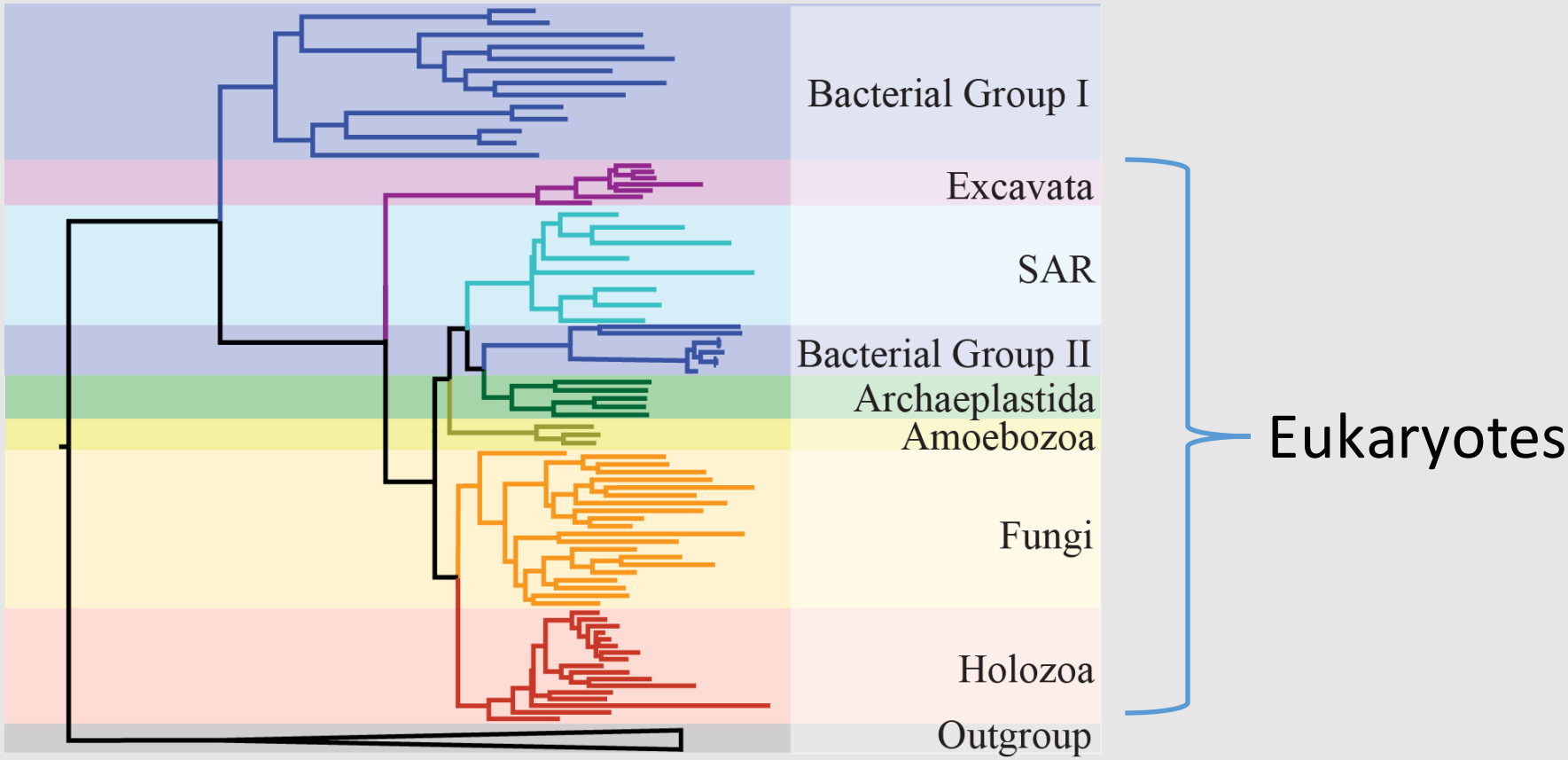
Wei JH, Yin X, Welander PV. Sterol synthesis in diverse bacteria. Frontiers in microbiology. 2016;7.

Methodology

- Collect SQMO and OSC gene from diverse taxa
- Reconstruct evolutionary trees of the SQMO and OSC genes
- Generate molecular clocks to determine when each gene separated from its last common ancestor (outgroup)

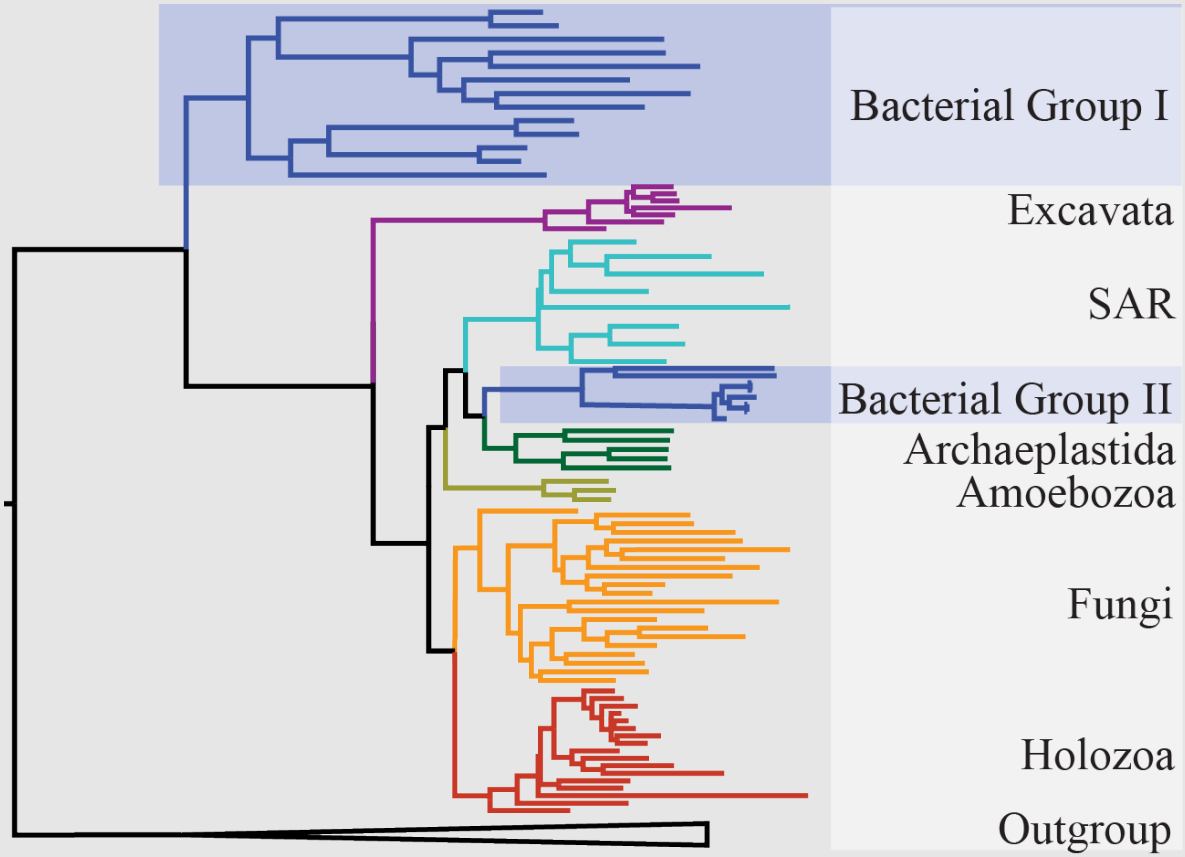


An evolutionary tree of SQMO



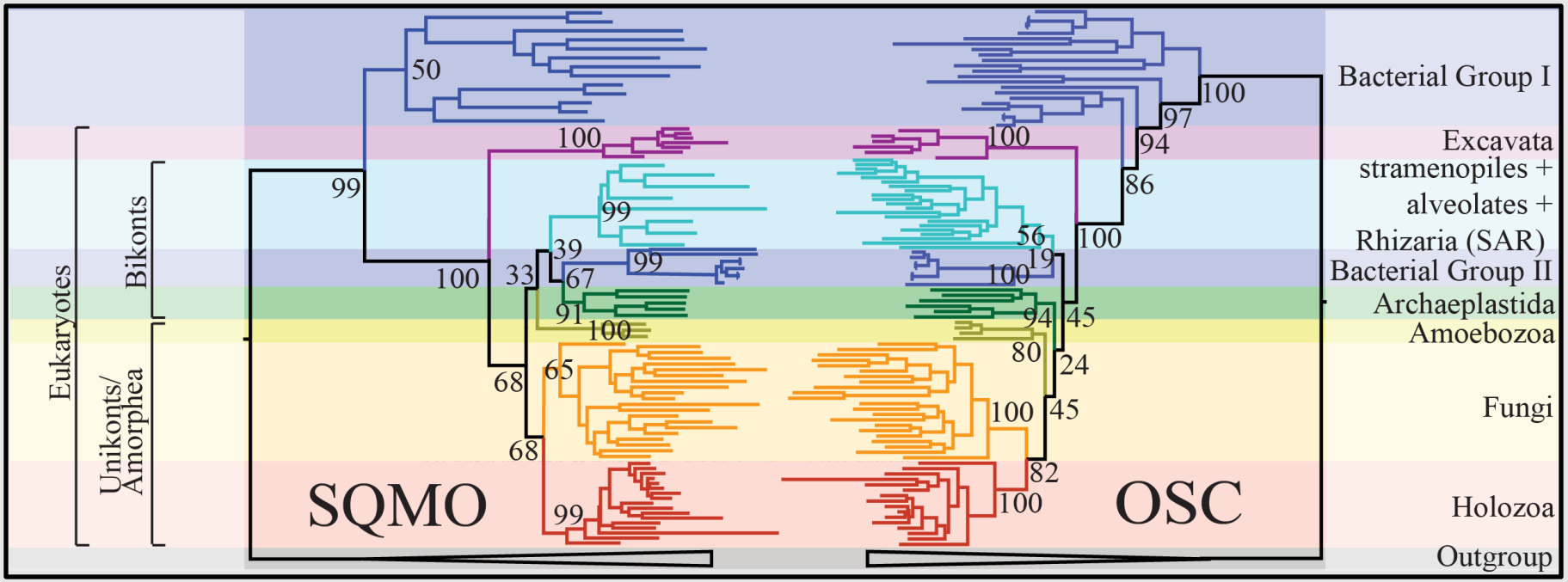
- Some bacterial sequences predate living (crown-group) eukaryotes

An evolutionary tree of SQMO



- Two different bacterial groups
- The members of these groups are not closely related
- This suggests two horizontal gene transfer events

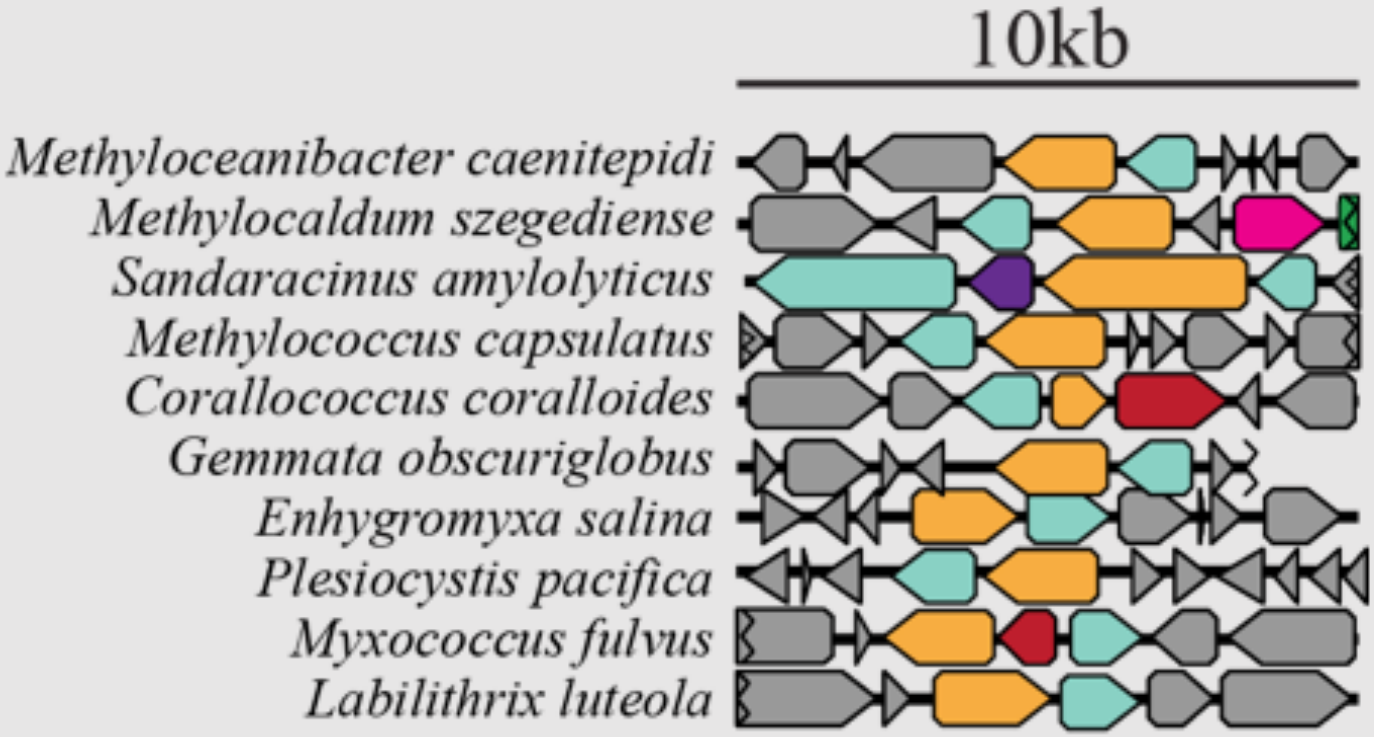
SQMO and OSC show the same evolutionary history



Perhaps the genes are being transferred as a pair?

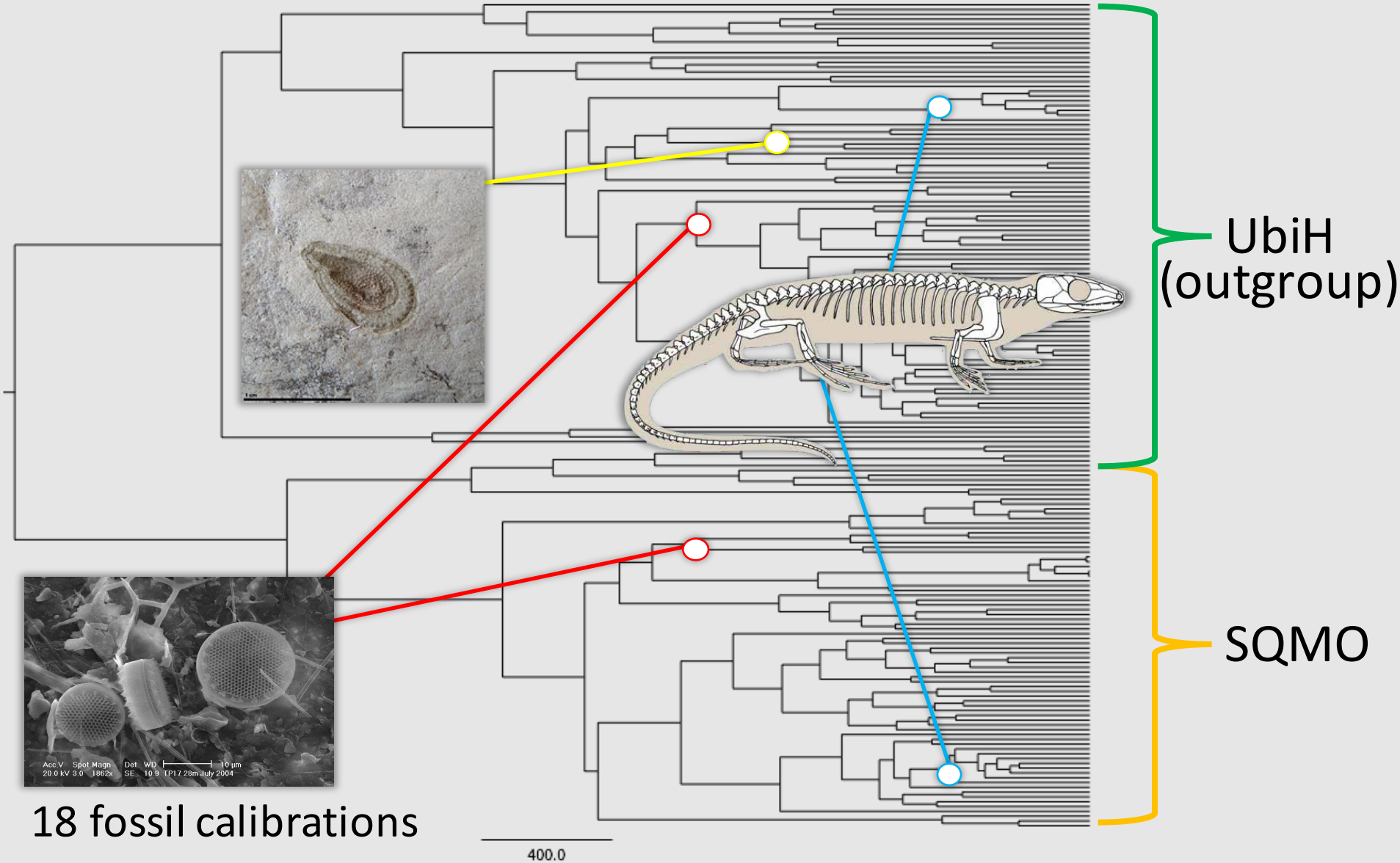
Gene synteny across bacterial genomes

Synteny: the physical co-localization of genes in the genome

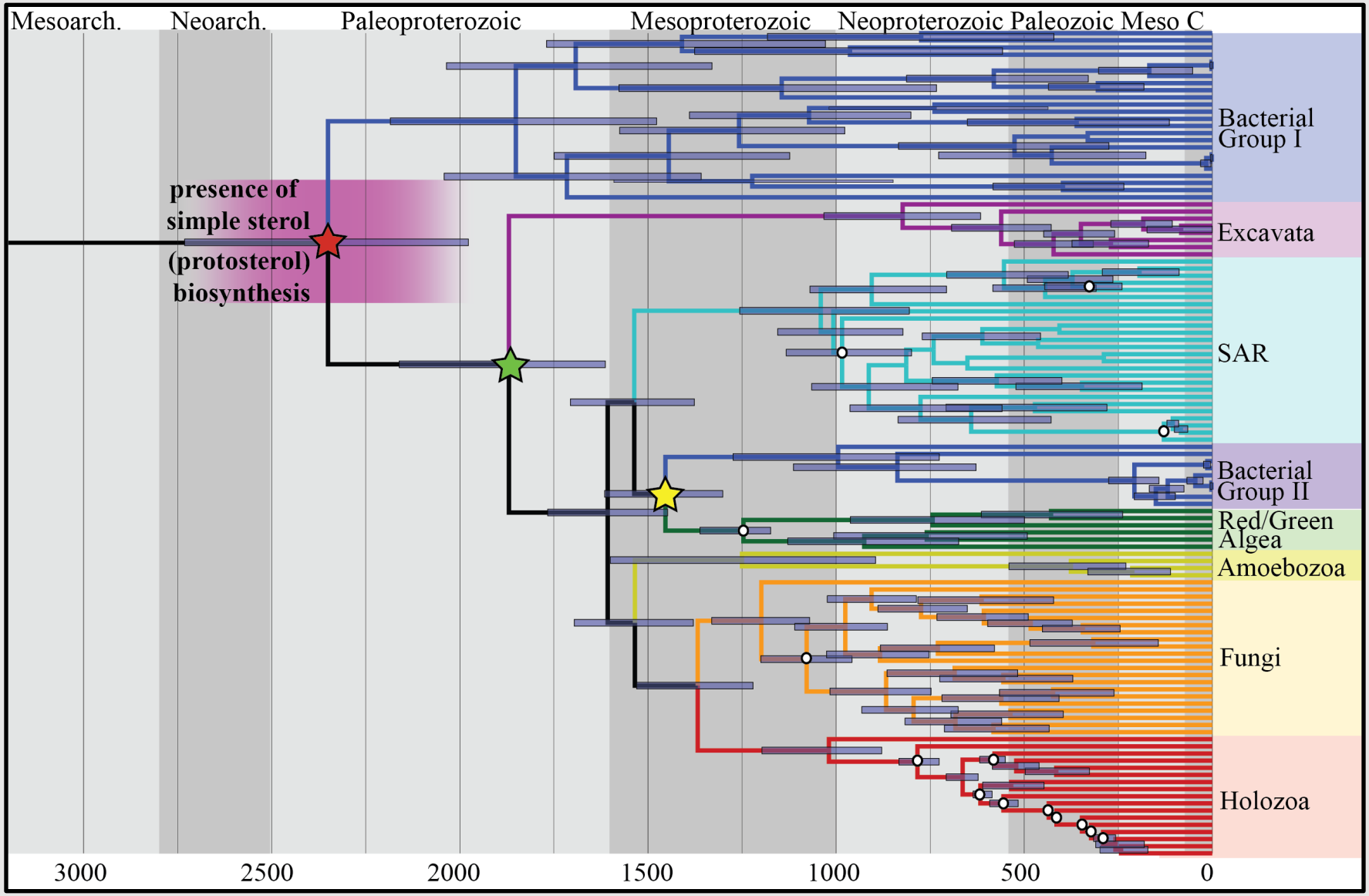


- The fact that both genes stay close together in such distantly related organisms supports horizontal gene transfer
- This means we can treat both genes like a single unit

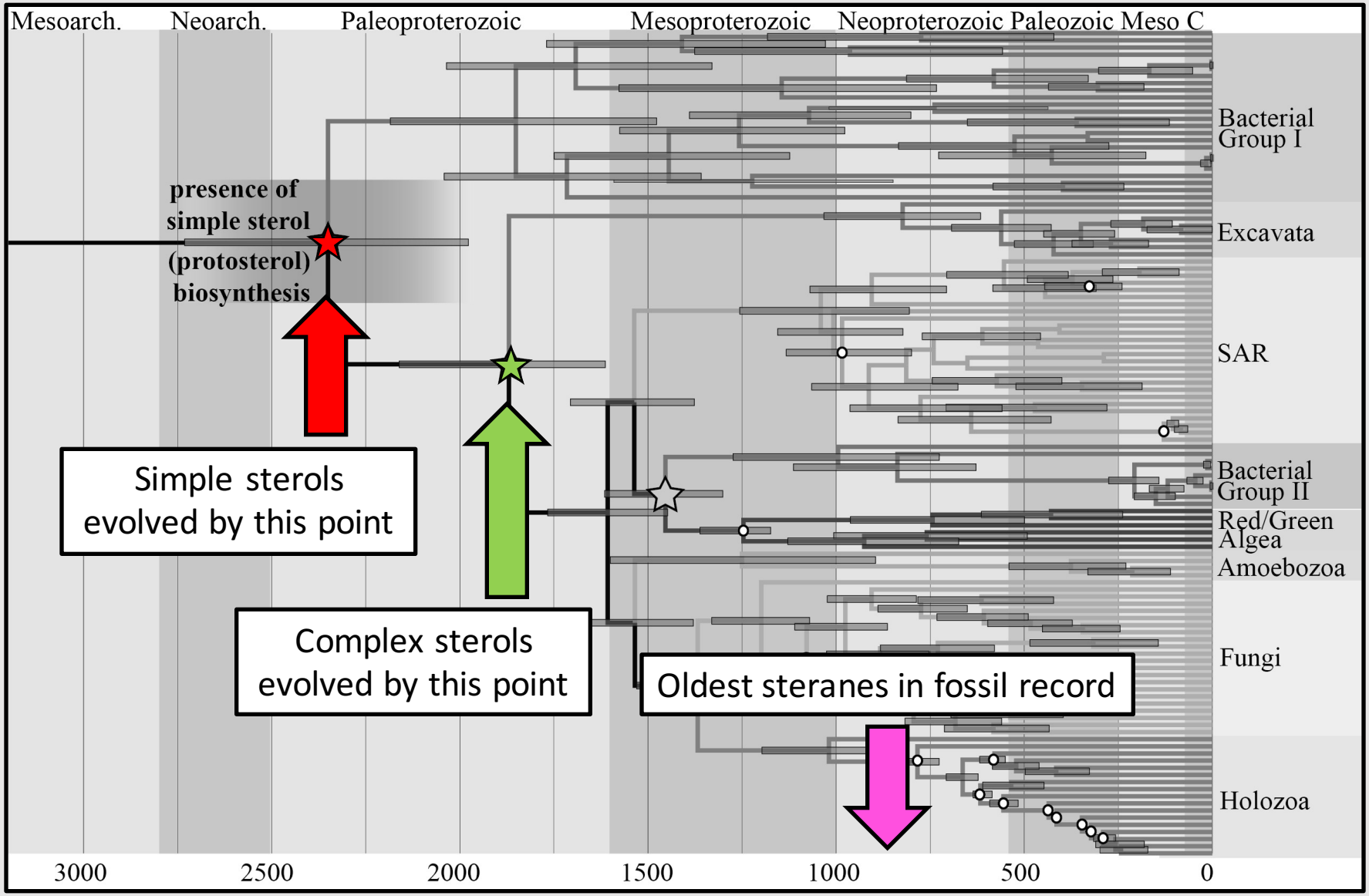
Creating the molecular clock



SQMO + OSC Molecular Clock



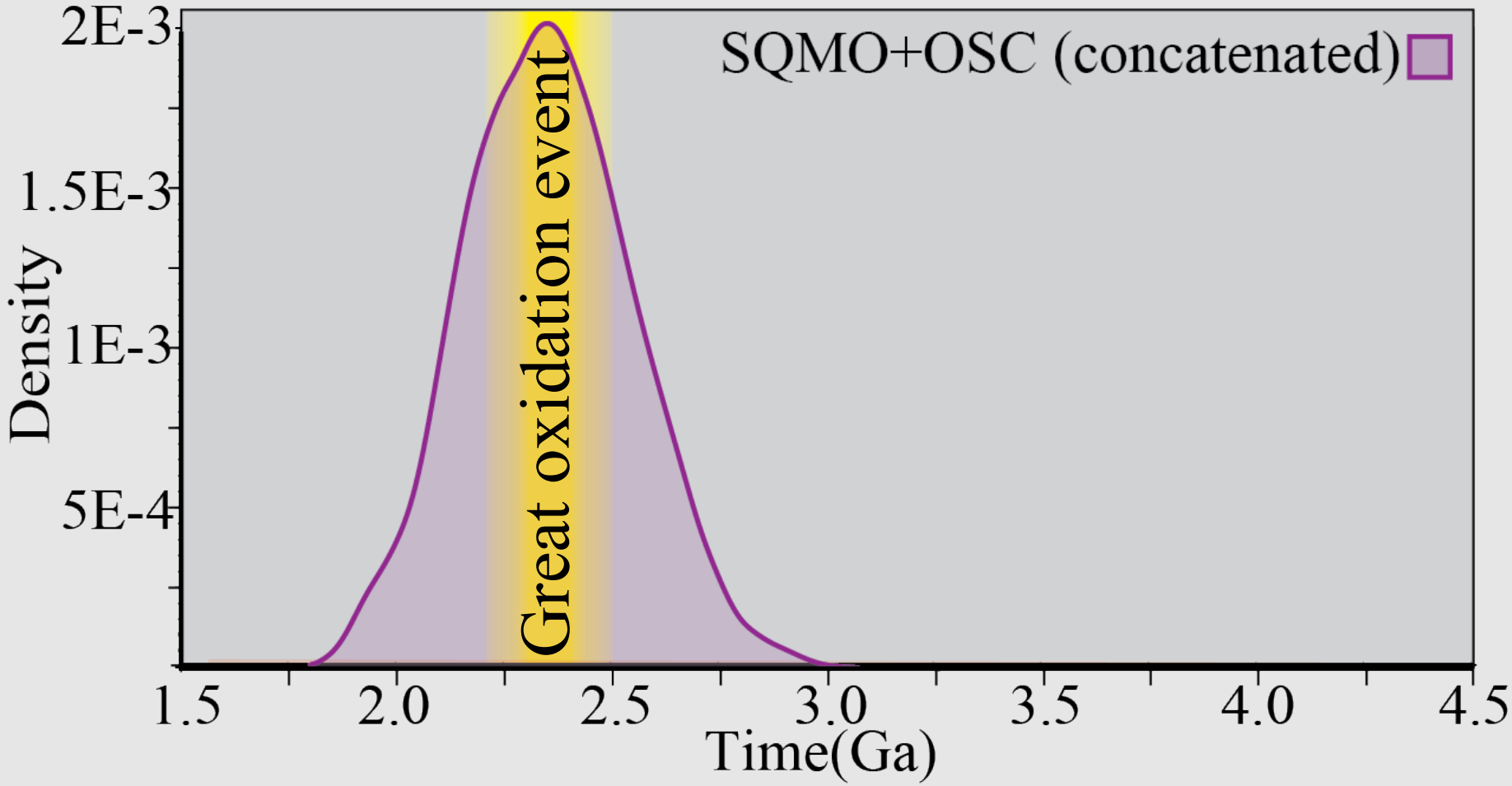
SQMO + OSC Molecular Clock



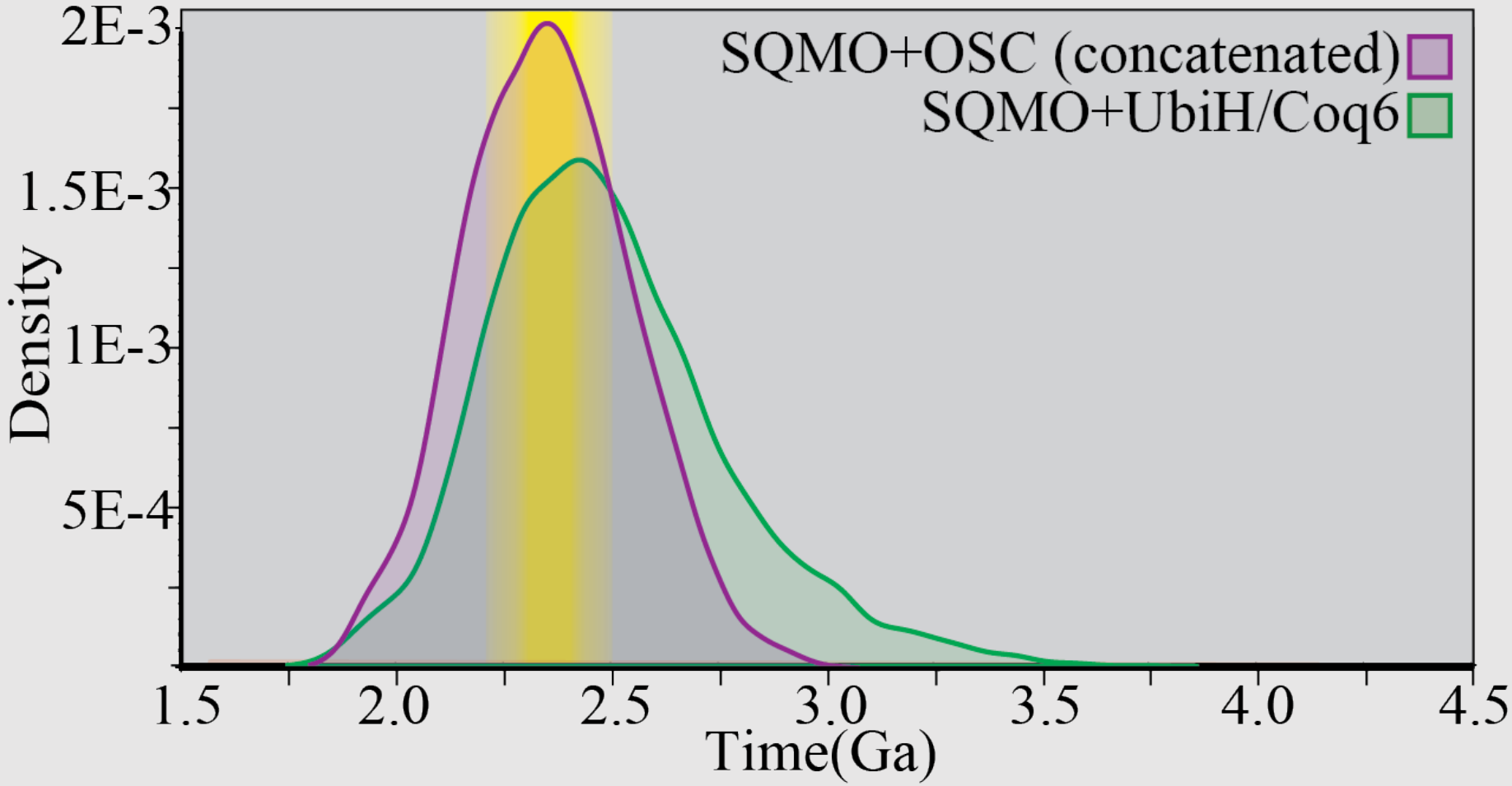
Caveat: there's a lot of uncertainty

Gene	Outrgoup?	Toplogy	Simple sterols evolved by this point	Complex sterols evolved by this point
SQMO	No	1	3.21(2.49-4.03)	2.22 (1.78-2.73)
SQMO	No	2	3.27 (2.51-4.18)	1.98 (1.58-2.34)
SQMO	Yes	1	2.47 (1.98-3.05)	1.80 (1.51-2.12)
SQMO	Yes	2	2.34(1.78-2.95)	1.60(1.40-1.92)
OSC	No	1	2.23(1.78-2.77)	1.61(1.39-1.87)
OSC	No	2	2.23(1.75-2.75)	1.51(1.30-1.81)
OSC	Yes	1	2.34(1.93-2.82)	1.56(1.37-1.78)
OSC	Yes	2	2.31(1.95-2.73)	1.51(1.33-1.70)
Both	No	1	2.28(1.90-2.71)	1.82(1.56-2.17)
Both	No	2	2.30(1.87-2.78)	1.60(1.45-1.80)

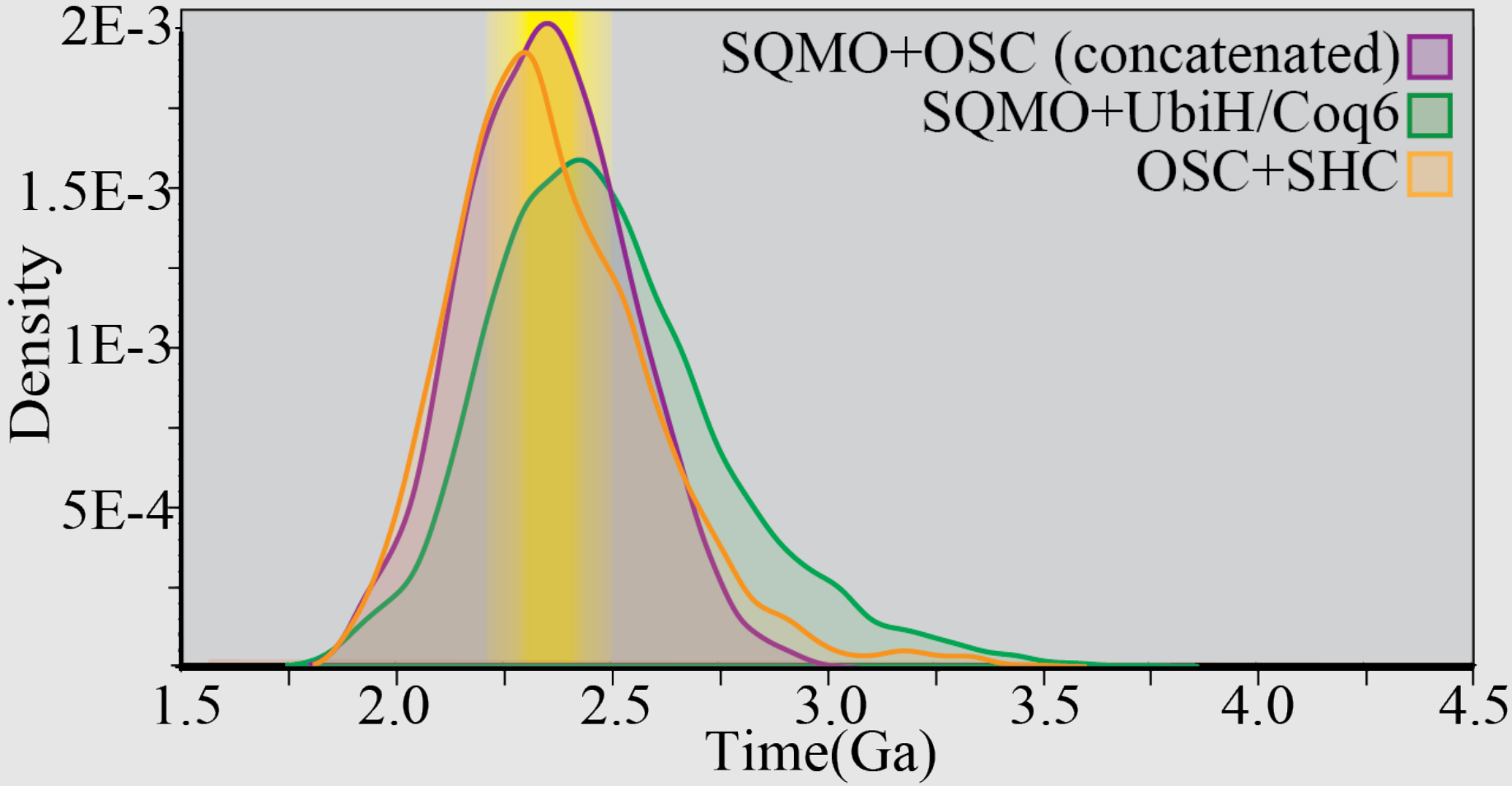
Visualizing uncertainty through the probability distribution



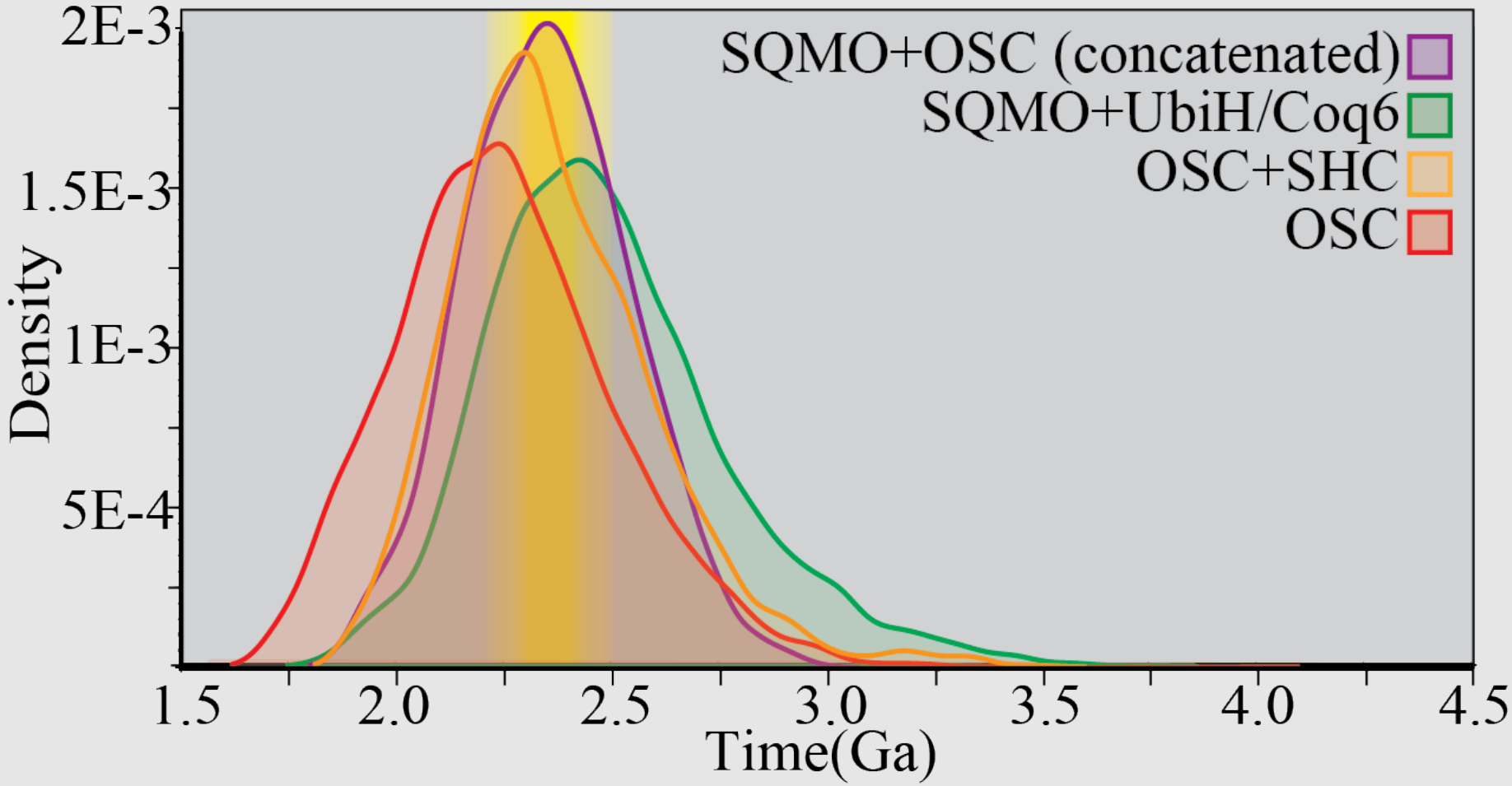
Visualizing uncertainty through the probability distribution



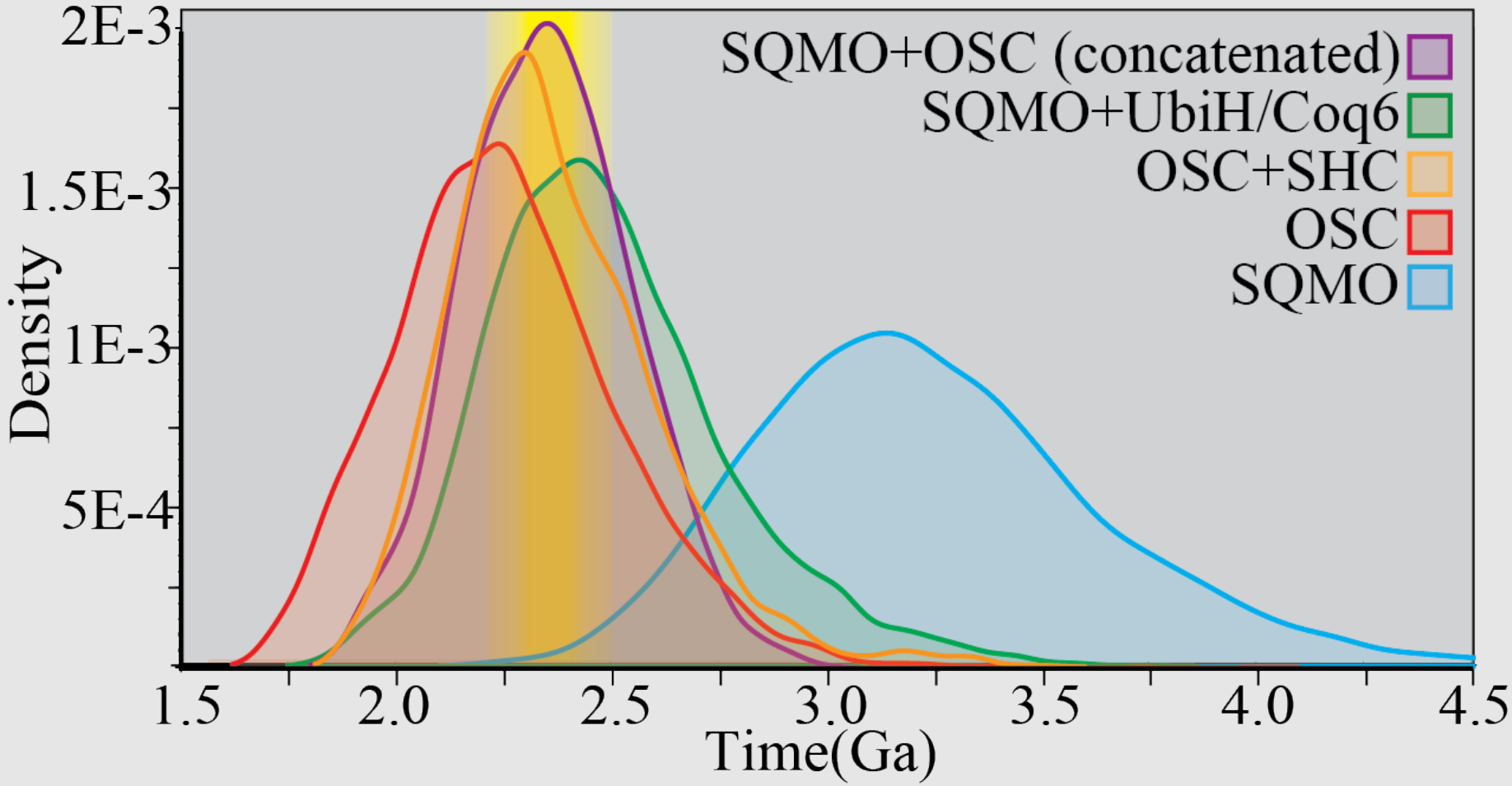
Visualizing uncertainty through the probability distribution



Visualizing uncertainty through the probability distribution



Visualizing uncertainty through the probability distribution



Conclusions

- Protosterol biosynthesis probably existed by the time oxygen is first detectable as a permanent presence in the Earth's atmosphere

- Sterol biosynthesis substantially predates the evolution of crown-group eukaryotes
- Our work suggests a >670 million year gap exists between the genetic and fossil records

How do we interpret this discrepancy?

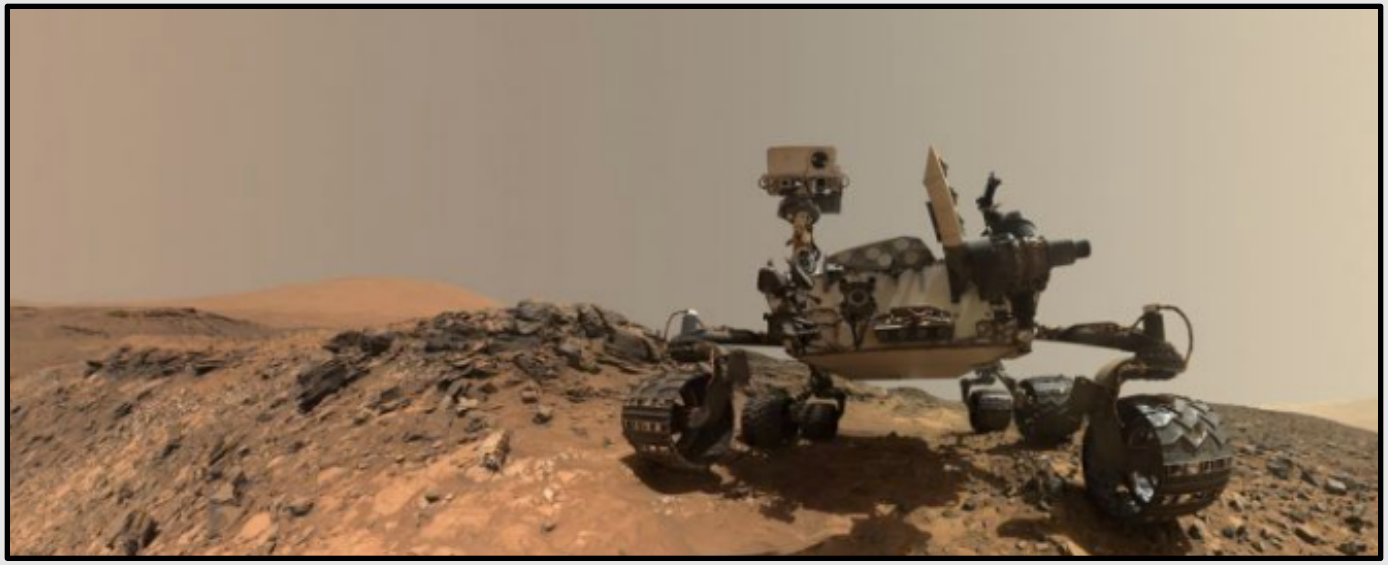
- Sponge biomarkers predate sponge fossils by ~50-100 My (Gold et al. 2016)
- Sponge molecular clocks predate sponge biomarkers by ~100 My (dos Reis et al. 2015)
- Algae molecular clocks predate fossils by ~400 My (Yang et al. 2016)

All of this data supports a "long fuse" hypothesis

Gold, David A., et al. "Sterol and genomic analyses validate the sponge biomarker hypothesis." Proceedings of the National Academy of Sciences 113.10 (2016): 2684-2689.dos
Reis, Mario, et al. "Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales." Current Biology 25.22 (2015): 2939-2950.
Yang, Eun Chan, et al. "Divergence time estimates and the evolution of major lineages in the florideophyte red algae." Scientific reports 6 (2016).

Implications for finding biosignatures on Earth and beyond

- The early history of many life forms involved a long period of low population
- It takes a long time for organisms to become widespread enough to leave a record in the geological data



- Some biochemical signals from complex life predate complex life itself

Acknowledgements

Abigail Caron
Greg Fournier

AGOURON INSTITUTE

THE SUMMONS LAB

FCL

NASA Astrobiology Institute
Foundation of Complex Life Team

NASA

DavidAdlerGold.com