A Timeline of Bacterial and Archaeal Diversification in the Ocean

Carolina A. Martinez-Gutierrez[1*], Josef C. Uyeda[1], Frank O. Aylward[1,2*]
1 Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA
2 Center for Emerging, Zoonotic, and Arthropod-borne Pathogens, Virginia Tech, Blacksburg, VA, USA

*Email for correspondence:
Frank O. Aylward: faylward@vt.edu; Carolina A. Martinez-Gutierrez: cmartinez@vt.edu

ABSTRACT

Microbial plankton play a central role in marine biogeochemical cycles, but the timing in which abundant lineages diversified into ocean environments remains unclear. Here, we reconstructed the timeline in which major clades of bacteria and archaea colonized the ocean using a high-resolution benchmarked phylogenetic tree that allows for simultaneous and direct comparison of the ages of multiple divergent lineages. Our findings show that the diversification of the most prevalent marine clades spans throughout a period of 2.2 Ga, with most clades colonizing the ocean during the last 800 million years. The oldest clades - SAR202, SAR324, *Ca.* Marinimicrobia, and Marine Group II - diversified around the time of the Great Oxidation Event (GOE), during which oxygen concentration increased but remained at microaerophilic levels throughout the Mid-Proterozoic, consistent with the prevalence of some clades within these groups in oxygen minimum zones today. We found the diversification of the prevalent heterotrophic marine clades SAR11, SAR116, SAR92, SAR86, and Roseobacter as well as the Marine Group I, to occur near to the Neoproterozoic Oxygenation Event (0.8-0.4 Ga). The diversification of these clades is concomitant with an overall increase of oxygen and nutrients in the ocean at this time, as well as the diversification of eukaryotic algae, consistent with the previous hypothesis that the diversification of heterotrophic bacteria is linked to the emergence of large eukaryotic

phytoplankton. The youngest clades correspond to the widespread phototrophic clades *Prochlorococcus, Synechococcus,* and *Crocosphaera*, whose diversification happened after the Phanerozoic Oxidation Event (0.45-0.4 Ga), in which oxygen concentrations had already reached their modern levels in the atmosphere and the ocean. Our work clarifies the timing at which abundant lineages of bacteria and archaea colonized the ocean, thereby providing key insights into the evolutionary history of lineages that comprise the majority of prokaryotic biomass in the modern ocean.

Keywords: Microbial Oceanography, Marine Microbial Diversification, Bayesian Molecular Dating, Great Oxidation Event

INTRODUCTION

The ocean plays a central role in the fluxes and stability of Earth's biogeochemistry (Dontsova et al., 2020; Falkowski et al., 1998; Field et al., 1998). Due to their abundance, diversity, and physiological versatility, microbes mediate the vast majority of organic matter transformations that underpin higher trophic levels (Brown et al., 2014; Mason et al., 2009). For example, marine microorganisms regulate a large fraction of the organic carbon pool (Ducklow and Doney, 2013), drive elemental cycling of nutrients like nitrogen (Zehr and Kudela, 2011), and participate in the ocean-atmosphere exchange of climatically important gasses (Vila-Costa et al., 2006). Starting in the 1980s, analysis of small-subunit ribosomal RNA genes began to reveal the identity of dominant clades of bacteria and archaea that were notable for their ubiquity and high abundance, and subsequent analyses highlighted their diverse physiological activities in the ocean (Giovannoni and Stingl, 2005). Phylogenetic studies showed that these clades are broadly distributed across the

Tree of Life (ToL) and encompass a wide range of phylogenetic breadths (Giovannoni and Stingl, 2005). Cultivation-based studies and the large-scale generation of genomes from metagenomes have continued to make major progress in examining the genomic diversity and metabolism of these major marine clades, but we still lack a comprehensive understanding of the evolutionary events leading to their origin and diversification in the ocean.

Several independent studies have used molecular phylogenetic methods to date the diversification of some marine microbial lineages, such as the Ammonia Oxidizing Archaea of the order *Nitrososphaerales* (Marine Group I, MGI) (Ren et al., 2019; Yang et al., 2021; Zhang et al., 2021), picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus* (Sánchez-Baracaldo, 2015; Sánchez-Baracaldo et al., 2019; Zhang et al., 2021), and marine alphaproteobacterial groups that included the SAR11 and Roseobacter clades (Luo et al., 2013). Differences in the methodological frameworks used in these studies often hinder comparisons between lineages, however, and results for individual clades often conflict (Ren et al., 2019; Sánchez-Baracaldo, 2015; Yang et al., 2021; Zhang et al., 2021). Moreover, it has been difficult to directly compare bacterial and archaeal clades due to the vast evolutionary distances between these domains. For these reasons it has remained challenging to evaluate the ages of different marine lineages and develop a comprehensive understanding of the timing of microbial diversification events in the ocean and their relationship with major geological events throughout Earth's history.

To clarify the timing at which major lineages of bacteria and archaea diversified into the ocean, we developed an approach that leverages a multi-domain phylogenetic tree that allows for simultaneous dating of all major marine lineages. This method allows us to directly compare the

ages of divergent lineages across the ToL and subsequently reconstruct a timeline in which these groups evolved into the ocean. Moreover, we can also map the acquisition of different protein families onto this phylogeny and thereby infer the genes that were gained by these marine lineages at the time of their emergence. Altogether, our study provides a comprehensive framework that sheds light on watershed events in the history of life on Earth that have given rise to contemporary biodiversity and biogeochemical dynamics in the ocean.

RESULTS AND DISCUSSION

To begin analyzing the diversification of marine lineages of bacteria and archaea, we constructed a multi-domain phylogenetic tree that allowed us to directly compare the origin of 13 planktonic marine bacterial and archaeal clades that are notable for their abundance and major roles in marine biogeochemical cycles (Fig. 1). We based tree reconstruction on a benchmarked set of marker genes that we have previously shown to be congruent for inter-domain phylogenetic reconstruction (Martinez-Gutierrez and Aylward, 2021) (details in Methods, Supplemental File 2). Our phylogenetic framework included non-marine clades for phylogenetic context, and overall it recapitulates known relationships across the ToL, such as the clear demarcation of the Gracilicutes and Terrabacteria bacterial superphyla and the basal placement of the *Thermatogales* within Bacteria (Coleman et al., 2021; Martinez-Gutierrez and Aylward, 2021) (Fig. 1). To gain insight into the geological landscape in which these major marine clades first diversified, we performed a Bayesian relaxed molecular dating analysis on our benchmarked ToL using several calibrated nodes (Fig. 1 and Table 1).

Due to the limited representation of microorganisms in the fossil record and the difficulties to associate fossils to extant relatives, we employed geochemical evidence as temporal calibrations (Fig. 1 and Table 1). Moreover, because of the uncertainty in the length of the branch linking bacteria and archaea, the crown node for each domain was calibrated independently. We used both the age of the presence of liquid water (as approximated through the dating of zircons (Valley et al., 2014)) as well as the most ancient record of biogenic methane (broadly used as evidence of life on Earth (Ueno et al., 2006)) as maximum and minimum prior ages for bacteria and archaea (4400 and 3460 My, respectively, Fig. 1 and Table 1). For internal calibration, we used the recent identification of non-oxygenic Cyanobacteria to constrain the diversification node of oxygenic Cyanobacteria with a minimum age of 2320 My, the estimated age for the Great Oxidation Event (GOE) (Bekker et al., 2004; Holland, 2006, 2002). Similarly, we calibrated the crown node of aerobic Ammonia Oxidizing Archaea, aerobic *Ca.* Marinimicrobia, and the Nitrite Oxidizing Bacteria with a maximum age of 2320 My (GOE estimated age) due to their strict aerobic metabolism. Despite geological evidence pointing to the presence of oxygen before the GOE, our Bayesian estimates indicate an overall consistency of the priors used (Supplemental File 5), and we recovered the ancient origin of major bacterial and archaeal supergroups, such as Asgardarchaeota, Euryarchaeota, Firmicutes, Actinobacteria, and Aquificota (Fig. 2). Moreover, the date we found for oxygenic Cyanobacteria (2611 My, CI 95% = 2589-2632; Fig. 2) is in agreement with their diversification happening before the GOE (Ward et al., 2016). Please see Table 1 for a detailed explanation of all calibration dates used, together with our rationale for including each one.

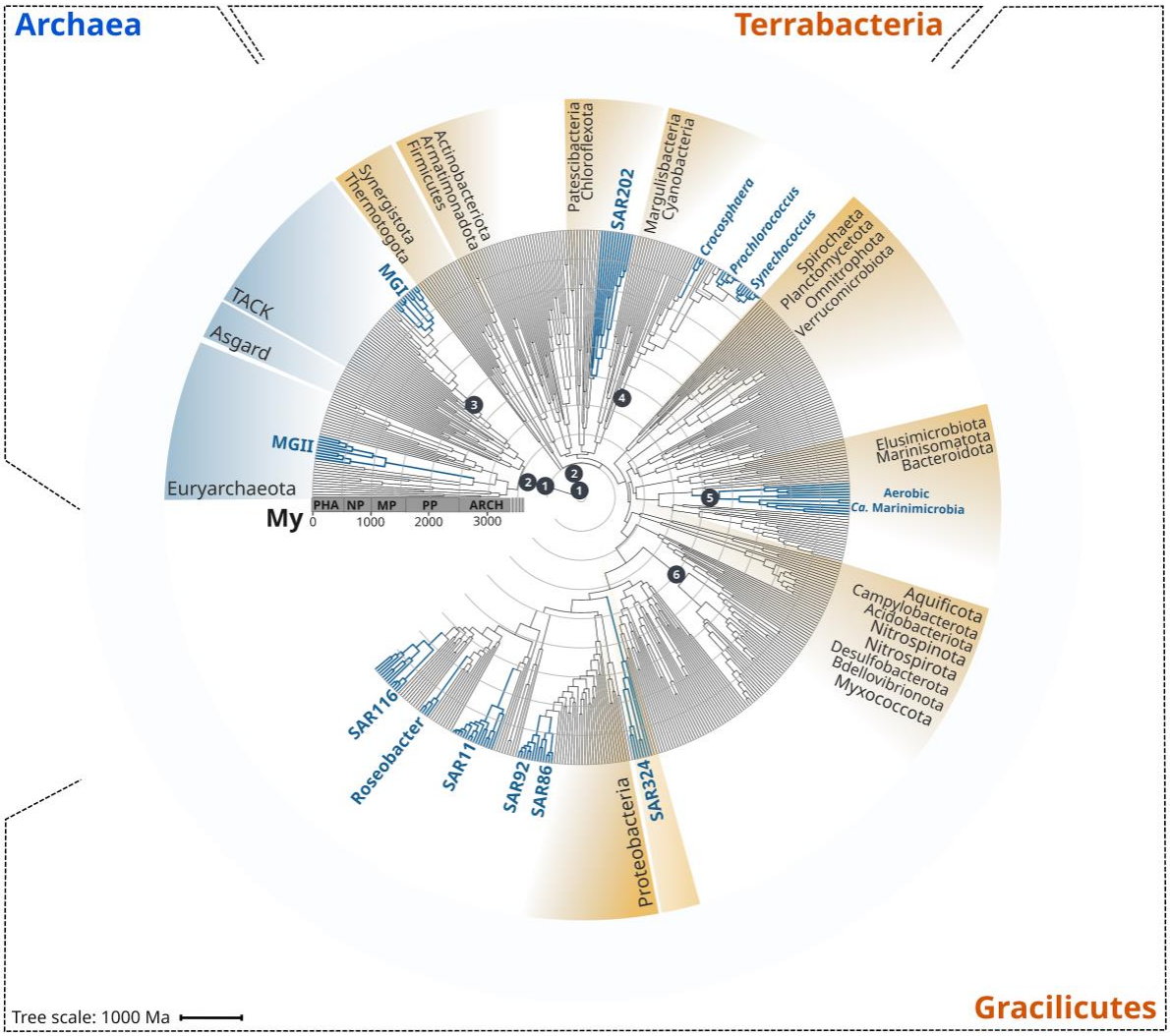

Figure 1. Rooted inter-domain Tree of Life used for molecular dating analyses. Maximum likelihood tree constructed with IQ-TREE v1.6.12 using the concatenation of 30 RNAP subunits and ribosomal protein sequences and the substitution model LG+R10. Blue labels represent the marine clades dated in our study. Dark gray dots show the temporal calibration used in our molecular dating analyses (Table 1). The marine clades shown are classified on the GTDB as follows: MGII, *Poseidoniales;* MGI, *Nitrososphaerales*; SAR202, SAR202; *Crocosphaera*, *Crocosphaera*; *Prochlorococcus*, *Prochlorococcus*; *Synechococcus*, *Synechococcus*; *Ca.* Marinimicrobia, *Marinisomatia*; SAR324, SAR324; SAR86, *Oceanospirillales*; SAR92, *Porticoccaceae*; SAR11, *Pelagibacterales*; Roseobacter, *Rhodobacteraceae*; SAR116, *Puniceispirillaceae.*

Table 1. Temporal calibrations used as priors for the molecular dating of the main marine microbial clades. See methods for a detailed explanation of the calibrations used.

Node	Calibration group	Minimum (MY)	Maximum (MY)	Evidence	Reference
1,2	Bacteria-Archaea Root	-	4400	Identification of the most ancient zircons showing evidence of liquid water.	(Valley et al., 2014)
1,2	Bacteria-Archaea Root	3460	-	Identification of the most ancient traces of methane. Minimum age for life on Earth. Calibration consistent with the most ancient fossils found to date (~3.5 Ga; Walter et al. 1980).	(Ueno et al., 2006)
3	Aerobic *Nitrososphaerales*	-	2320	Strict aerobic metabolism.	(Ueno et al., 2006)
4	Oxygenic Cyanobacteria	2320	-	Oxygenation of the atmosphere. The Great Oxidation Event has been associated with oxygenic Cyanobacteria.	(Bekker et al., 2004; Holland, 2006, 2002)
5	Aerobic *Ca.* Marinimicrobia	-	2320	Strict aerobic metabolism.	(Bekker et al., 2004; Holland, 2006, 2002)
6	Nitrite oxidizing bacteria	-	2320	Strict aerobic metabolism.	(Bekker et al., 2004; Holland, 2006, 2002)

Our Bayesian estimates suggest that the lineages that emerged earliest are SAR202, aerobic *Ca.* Marinimicrobia, SAR324, and the Marine Group II of the phylum Euryarchaeota (MGII). The most ancient clade was the SAR202 (2479 My, 95% CI = 2465-2492 My), whose diversification took place near before the GOE (Fig. 2). Given the broadly distributed aerobic capabilities of SAR202, the diversification of this clade before the GOE suggests that SAR202 emerged during an oxygen oasis proposed to have existed in pre-GOE Earth (Anbar et al., 2007; Ossa Ossa et al., 2019; Reinhard and Planavsky, 2022). The ancient pre-GOE origin of SAR202 is consistent with a recent study that proposed that this clade played a role in the shift of the redox state of the

atmosphere during the GOE. SAR202 is able to partially metabolize organic matter through a flavin dependent Baeyer–Villiger monooxygenase, thereby enhancing the burial of organic matter and contributing to the net accumulation of oxygen in the atmosphere (Landry et al., 2017; Shang et al., 2022). After the GOE, we detected the diversification of aerobic *Ca.* Marinimicrobia (2196 My, 95% CI = 2173-2219 My), the SAR324 clade (1686 My, 95% CI = 1658-1715 My), and the MGII clade (1184 My, 95% CI = 1166-1202 My) (Fig. 2). Although these ancient clades may have first diversified under the oxic conditions derived from the GOE, it has been suggested that the initial oxygenation of Earth was followed by a relatively rapid drop in ocean and atmosphere oxygen levels (Alcott et al., 2019; Hodgskiss et al., 2019; Reinhard and Planavsky, 2022). It is therefore likely that these clades diversified in the microaerophilic and variable oxygen conditions that prevailed during this period (Bekker et al., 2004; Holland, 2006, 2002). Indeed, the oxygen landscape in which these marine clades first diversified is consistent with their current physiology. For example, these groups are capable of using oxygen as well as alternative electron acceptors (e.g., nitrate and sulfate), and several representatives are prevalent in modern marine oxygen minimum (OMZs) (Pajares et al., 2020; Sheik et al., 2014; Thrash et al., 2017; Ulloa et al., 2012). The facultative aerobic or microaerophilic metabolism in these clades is therefore potentially a vestige of the low oxygen environment of most of the Proterozoic Eon, and in this way OMZs can be considered to be modern-day refugia of these ancient ocean conditions. Of the clades that diversified as part of this early period, MGII and SAR324 show the youngest colonization dates, but we suspect that this may be due to the notably long branches that lead to the crown nodes of these lineages. These long branches are likely caused by the absence of basal-branching members of these clades ─ either due to extinction events or under-sampling of rare lineages in the available genome collection ─ that would have increased the age of these lineages if present in the tree.

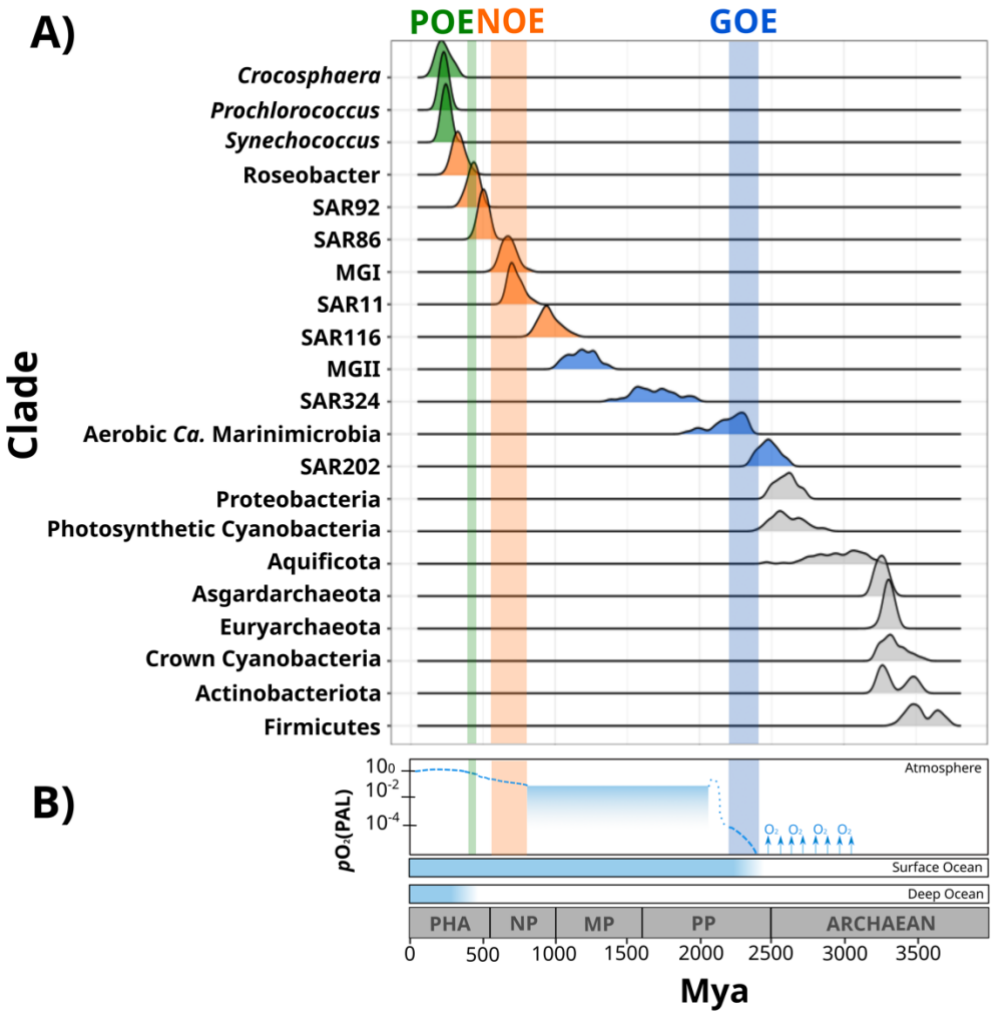

Figure 2. Dates of the diversification of marine microbial clades and their relationship with the redox history of Earth's atmosphere, surface ocean, and deep ocean. A) Ridges represent the distribution of 100 Bayesian dates estimated using a relaxed molecular clock and an autocorrelated model (see Methods). Ridges of marine clades were colored based on their diversification date: green, Late-branching Phototrophs; orange, Late-branching Clades; blue, Early-branching Clades. The timing of the diversification of major bacterial and archaeal superphyla is represented with gray ridges. Molecular dating estimates resulting from the uncorrelated model UGAM and the autocorrelated model CIR are shown on Supplemental File 8. B) Oxygenation events and redox changes across Earth's history. Panel adapted from previous work (Alcott et al., 2019). Abbreviations: POE, Paleozoic Oxidation Event; NOE, Neoproterozoic Oxidation Event; GOE, Great Oxidation Event; Pha, Paleozoic; NP, Neoproterozoic; MP: Mesoproterozoic; PP: Paleoproterozoic.

According to our analysis, the next clades to diversify in the ocean are SAR116 (959 My, 95% CI = 945-973 My), SAR11 (725 My, 95% CI = 715-734 My), SAR86 (503 My, 95% CI = 497-509 My), SAR92 (430 My, 95% CI = 423-437 My), and Roseobacter (332 My, 95% CI = 323-340 My) (Fig. 2). The relatively late appearance of these heterotrophic lineages that are abundant in the open ocean today was potentially due to the low productivity and oxygen concentrations in both shallow and deep waters that prevailed in the Mid-Proterozoic (1800-800 My), a period previously described as the "boring billion" (Anbar and Knoll, 2002; Crockford et al., 2018; Hodgskiss et al., 2019; Planavsky et al., 2014; Tang et al., 2016). The diversification of these clades may be indirectly associated with the Snowball event registered before the Neoproterozoic Oxidation Event (NOE, 800-540 My) (Anbar and Knoll, 2002; Hoffman et al., 1998; Shields-Zhou and Och, 2011), which increased the availability of oxygen and inorganic nutrients in the ocean (Anbar and Knoll, 2002; Butterfield, 2001; Porter, 2004; Shields-Zhou and Och, 2011; Vidal and Moczydłowska-Vidal, 1997), and is also coincident with the widespread diversification of large eukaryotic algae during the Neoproterozoic (Anbar and Knoll, 2002; Butterfield, 2001; Porter, 2004; Shields-Zhou and Och, 2011; Vidal and Moczydłowska-Vidal, 1997) (Fig. 4). It is therefore plausible that an increase in nutrients as well as the broad diversification of eukaryotic plankton enhanced the mobility of organic and inorganic nutrients beyond the coastal areas, and increased the burial of organic matter that consequently led to the rise in atmospheric oxygen concentrations (Knoll et al., 2006; Shields-Zhou and Och, 2011). The scenario in which heterotrophic marine clades diversified in part as a consequence of the new niches built by marine eukaryotes has been previously proposed to have driven the diversification of the Roseobacter clade (Luo et al., 2013; Luo and Moran, 2014). The diversification timing of Roseobacter and other heterotrophic clades support this phenomenon and suggest that the interaction with marine eukaryotes may have

broadly influenced the diversification of prevalent lineages in the modern ocean. Similar to what we observed in MGII and SAR324, the Roseobacter clade shows a long branch leading to the crown node (Fig. 1), suggesting that the diversification of this clade may have occurred earlier.

We also report the diversification of the chemolithoautotroph archaeal lineage MGI into the ocean after the NOE (678 My, 95% CI = 668-688 My) (Fig 2), which is comparable with the age reported by another independent study (Yang et al., 2021). This is consistent with an increase in the oxygen concentrations of the ocean during this period (Reinhard and Planavsky, 2022), a necessary requisite for ammonia oxidation. Moreover, the widespread sulfidic conditions that likely prevailed during the Mid-Proterozoic ocean may have limited the availability of redox-sensitive metals such as copper, which is necessary for ammonia monooxygenases (Anbar and Knoll, 2002; Hatzenpichler, 2012). It is therefore possible that a low concentration of oxygen and limited inorganic nutrient availability before the NOE were limiting factors that delayed the colonization of AOA into the ocean.

The most recent lineages to emerge include the genera *Synechococcus* (243 My, 95% CI = 238-247 My), *Prochlorococcus* (230 My, 95% CI = 225-234 My), and the diazotroph *Crocosphaera* (228 My, 95% CI = 218-237 My). Our results agree with an independent study that points to a relatively late evolution of the marine cyanobacterial clades *Prochlorococcus* and *Synechococcus* (Sánchez-Baracaldo, 2015). Picocyanobacteria and *Crocosphaera* are critical components of phytoplanktonic communities in the modern open ocean due to their large contribution to carbon and nitrogen fixation, respectively (Flombaum et al., 2013; Montoya et al., 2004; Scanlan et al., 2009). For example, the nitrogen fixation activities of *Crocosphaera watsonii* in the open ocean today support the demands of nitrogen-starved microbial food webs found in oligotrophic waters

(Hewson et al., 2009). The relatively late diversification of these lineages suggests that the oligotrophic open ocean is a relatively modern ecosystem. Moreover, the oligotrophic ocean today is characterized by the rapid turnover of nutrients that depends on the efficient mobilization of essential elements through the ocean (Karl, 2002). Due to its distance from terrestrial nutrient inputs, productivity in the open ocean is therefore dependent on local nitrogen fixation, which was likely enhanced after the widespread oxygenation of the ocean that made molybdenum widely available due to its high solubility in oxic seawater (Canfield et al., 2007; Scott et al., 2008; Wei et al., 2021). Such widespread oxygenation was registered 430-390 My in an event referred to here as the Paleozoic Oxidation Event (Berner and Raiswell, 1983; Lenton et al., 2016; Sperling et al., 2015; Tostevin and Mills, 2020) (POE, Fig. 2 and 4). The increase of oxygen to present-day levels in the atmosphere and the ocean was potentially the result of an increment of the burial of organic carbon in sedimentary rocks due to the diversification of the earliest land plants (Lenton et al., 2016; Planavsky et al., 2021; Reinhard and Planavsky, 2022). The POE has also been associated with increased phosphorus weathering rates (Bergman et al., 2004; Lenton et al., 2016), global impacts on the global element cycles (Dahl and Arens, 2020), and an increase in overall ocean productivity (Planavsky et al., 2021). The late diversification of oligotrophic-specialized clades after the POE therefore suggests that the establishment of the oligotrophic open ocean as we know it today would only have been plausible once modern oxygen concentrations and biogeochemical dynamics were reached (Karl, 2002; Reinhard and Planavsky, 2022).

To shed further light on the drivers of the colonization of the ocean, we investigated whether the diversification of marine microbial clades was linked to the acquisition of novel metabolic capabilities. We broadly classified the different clades as Early-branching Clades, Late-branching

Clades, and Late-branching Phototrophs based on the general timing of their diversification (Fig. 2). To identify the enrichment of gene functions at the crown node of each marine clade (Fig. 1), we performed a stochastic mapping analysis on each of the 112,248 protein families encoded in our genome dataset. We compared our results with a null hypothesis distribution in which a constant rate model was implemented unconditionally of observed data (see Methods). Statistical comparisons of the stochastic and the null distribution show that each diversification phase was associated with the enrichment of specific functional categories that were consistent with the geochemical context of their diversification (Fig. 3 and 4). For example, Early-branching Clades (EBC) gained a disproportionate number of genes involved in DNA repair, recombination, and glutathione metabolism, consistent with the hypothesis that the GOE led to a rise in reactive oxygen species that cause DNA damage (Burrows, 2010; Khademian and Imlay, 2021; Masip et al., 2006). Moreover, the EBC were enriched in proteins involved in ancient aerobic pathways, such as oxidative phosphorylation and the TCA cycle (Fig. 3), as well as genes implicated in the degradation of fatty acids under aerobic conditions, such as the enzyme alkane 1-monooxygenase in MGII (Supplemental File 6). We also detected genes for the adaptation to marine environments, including genes for the anabolism of taurine (e.g., cysteine dioxygenase in MGII, Supplemental File 6), an osmoprotectant commonly found in marine bacteria (McParland et al., 2021). Our findings suggest that the diversification of EBC in the ocean was linked to the emergence of aerobic metabolism, the acquisition of metabolic capabilities to exploit the newly created niches that followed the increase of oxygen, and the expansion of genes involved in the tolerance to oxidative and salinity stress.

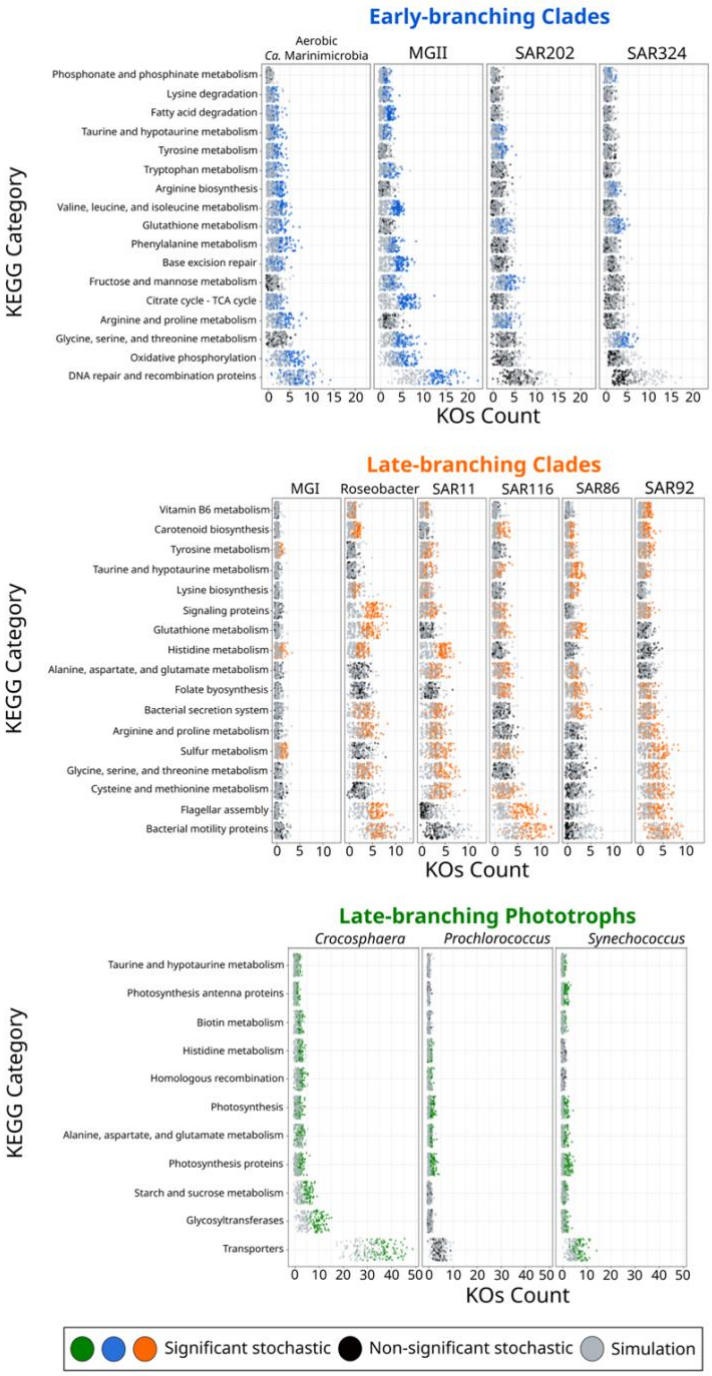

Figure 3. KEGG categories enriched at the crown node of each marine microbial clade. Clades were classified based on their diversification timing shown in Fig. 2. Enriched categories were identified by statistically comparing a stochastic mapping distribution with an all-rates-different model vs a null distribution with a constant rate model without conditioning on the presence/absence data at the tips. Each dot represents one replicate (See Methods). X-axis represents the number of KOs gained at each crown node for each KEGG category. Stochastic

mapping and null distributions were sorted for visualization purposes. The complete list of enriched KEGGs is shown in Supplemental File 7.

The emergence of Late-branching Clades (LBC; Fig. 3 and 4), whose diversification occurred around the time of the NOE and the initial diversification of eukaryotic algae (Parfrey et al., 2011), was characterized by the enrichment of substantially different gene repertories compared to EBC (Fig. 3). For instance, the heterotrophic lineages Roseobacter, SAR116, and SAR92 show an enrichment of flagellar assembly and motility genes (Fig. 3), including genes for flagellar biosynthesis, flagellin, and the flagellar basal-body assembly (Supplemental File 6). Motile marine heterotrophs like Roseobacter species have been associated with the marine phycosphere, a region surrounding individual phytoplankton cells releasing carbon-rich nutrients (Mühlenbruch et al., 2018; Seymour et al., 2017). Although the phycosphere can also be established between prokaryotic phytoplankton and heterotrophs (Croft et al., 2005; Seymour et al., 2017), given the late diversification of abundant marine prokaryotic phytoplankton (Fig. 2 and 4), it is plausible that the emergence of these clades was closely related to the establishment of ecological proximity with large eukaryotic algae. The potential diversification of heterotrophic LBC due to their ecological interactions with eukaryotic algae is further supported by the enrichment of genes involved in vitamin B6 metabolism and folate biosynthesis, which are key nutrients involved in phytoplankton-bacteria associations (Seymour et al., 2017). LBC were also enriched in genes for the catabolism of taurine (e.g., taurine transport system permease in SAR11 and a taurine dioxygenase in SAR86 and SAR92), suggesting that LBC gained metabolic capabilities to utilize the taurine produced by other organisms as a substrate (Clifford et al., 2019), instead of producing it as an osmoprotectant. Furthermore, we identified the enrichment of genes involved in carotenoid biosynthesis, including spheroidene monooxygenase, carotenoid 1,2-hydratase, beta-carotene

hydroxylase, and lycopene beta-cyclase (Supplemental File 6). The production of carotenoids is consistent with their use in proteorhodopsin, a light driven proton pump that is a hallmark feature of most marine heterotrophic bacteria, in particular those that inhabit energy-depleted areas of the ocean today (de la Torre et al., 2003).

Late-branching phototrophs that diversified around the time of the POE (LBP; Fig. 2) showed a remarkable enrichment of transporters in *Crocosphaera* and *Synechococcus* (Fig. 3). In particular, the diversification of *Crocosphaera* was characterized by the acquisition of transporters for inorganic nutrients like cobalt, nickel, iron, phosphonate, phosphate, ammonium, and magnesium, along with organic nutrients including amino acids and polysaccharides (Supplemental File 6). The acquisition of a wide diversity of transporters by the *Crocosphaera* is consistent with their boom-and-bust lifestyle seen in the oligotrophic open ocean today (Hewson et al., 2009; Wilson et al., 2017), which requires a rapid and efficient use of scarce nutrients. We also identified genes involved in osmotic pressure tolerance, for example a Ca-activated chloride channel homolog, a magnesium exporter, and a fluoride exporter (Supplemental File 6). In contrast, our results show that *Synechococcus* only acquired transporters for inorganic nutrients (e.g., iron and sulfate, Supplemental File 6), whereas *Prochlorococcus* did not show an enrichment of transporters (Supplemental File 6). Similar to LBC, we identified the enrichment of taurine metabolism genes in *Crocosphaera* and *Synechococcus*, suggesting that its use as osmoprotectant and potential substrate is widespread among planktonic microorganisms (Clifford et al., 2019). *Prochlorococcus* exhibits enrichment in fewer categories than the rest of phototrophic clades diversifying during the same period, consistent with the streamlined nature of genomes from this lineage (Partensky and Garczarek, 2010). The genes acquired by *Prochlorococcus* are involved in photosynthesis, which

supports previous findings that the diversification of this clade was accompanied by changes in the photosynthetic apparatus compared with *Synechococcus*, its sister group (Biller et al., 2015). Overall, the diversification of LBP was marked by the capacity to thrive in the oligotrophic ocean by exploiting organic and inorganic nutrients and modifying the photosynthetic apparatus as observed in *Crocosphaera* and *Synechococcus*, and *Prochlorococcus*, respectively.

CONCLUSION

The contemporary ocean is dominated by abundant clades of bacteria and archaea that drive global biogeochemical cycles and play a central role in shaping the redox state of the planet. Despite their importance, the timing and geological landscape in which these clades colonized the ocean have remained unclear due to a combination of the inherent difficulties of studying biological events that occurred in deep time and the lack of a fossil record for microbial life. Yet establishing a timeline of these events is critical because the colonization of major marine lineages led to the establishment of the biogeochemical cycles that govern the environmental health of the planet today. In this study, we develop a novel phylogenomic method that allows us to infer a comprehensive timeline of the colonization of the ocean by abundant marine clades of both bacteria and archaea. Importantly, our study presents key foundational knowledge for understanding ongoing anthropogenic changes in the ocean. Climate change is predicted to lead to an expansion of both oxygen minimum zones, which our findings suggest are refugia that date back to the mid-Proterozoic ocean, and oligotrophic surface waters, which represent ecosystems that emerged relatively recently in the Phanerozoic (Fig. 4). Thus, the impacts of current global change can manifest similarly in ecosystems that have emerged at dramatically different periods of Earth's history. Knowledge of how and under what geochemical conditions dominant microbial

constituents first diversified provides context for understanding the impact of climatic changes on the marine biome more broadly and will help to clarify how continuing ecological shifts will impact marine biogeochemical cycles.

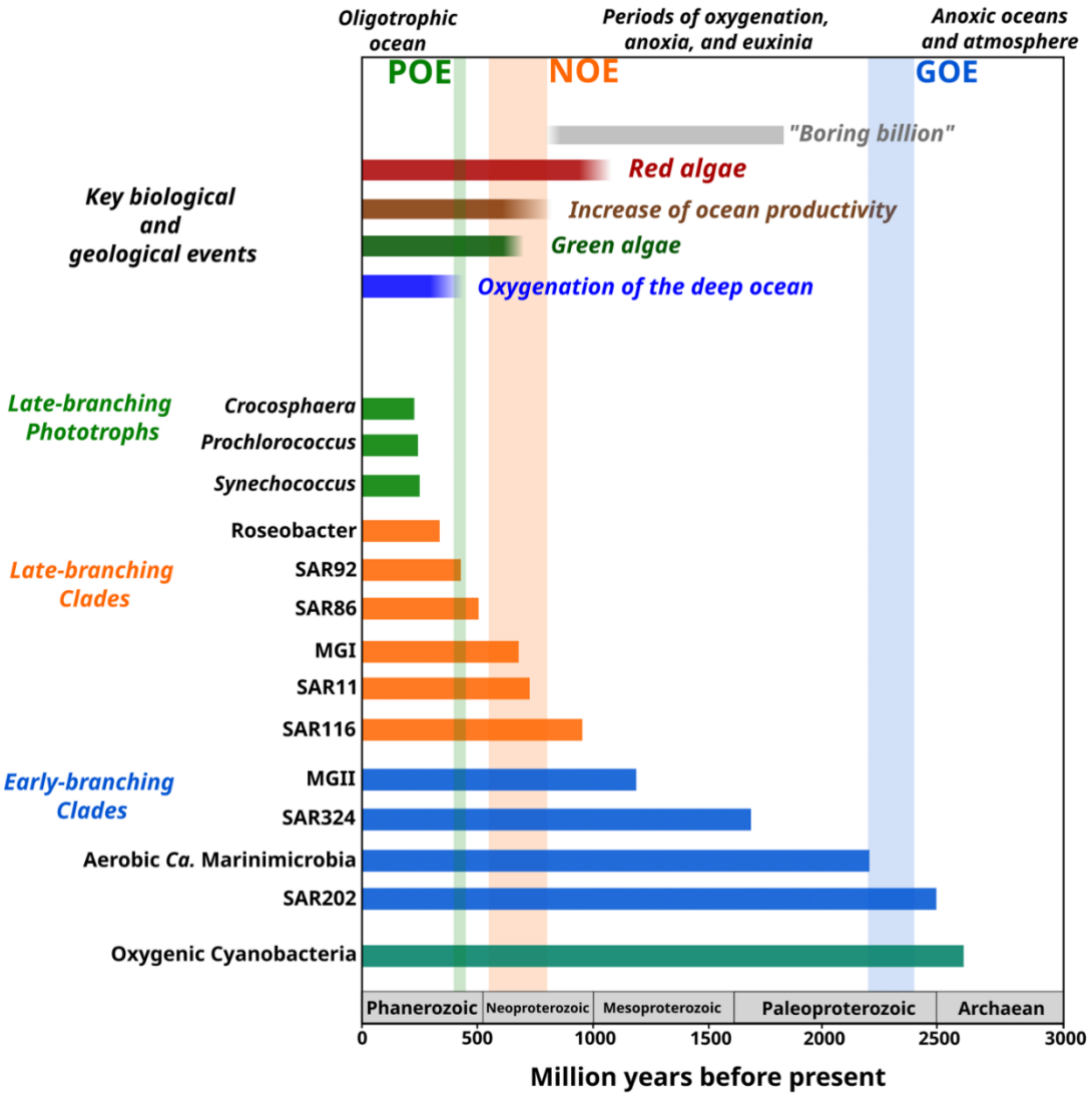

Figure 4. Link between the timing of the diversification of the main marine microbial clades and major geological and biological events. The timing of the geological and biological events potentially involved in the diversification of marine clades is based on previously published data: "Boring billion" (Brasier and Lindsay, 1998; Hodgskiss et al., 2019), red algae fossils (Butterfield, 2000), increased of ocean productivity (Butterfield, 2000; Och and Shields-Zhou, 2012), green algae fossils (Butterfield et al., 2006), and oxygenation of the deep ocean (Lenton et al., 2016). The length of each bar represents the estimated age for marine clades according to Bayesian estimates. The timing of the main oxygenation events is based on previous work (Alcott et al., 2019).

MATERIAL AND METHODS

Genome sampling and phylogenetic reconstruction

To obtain a comprehensive understanding of the diversification of the main marine planktonic clades, we built a multi-domain phylogenetic tree that included a broad diversity of bacterial and archaeal genomes. We compiled a balanced genome dataset from the Genome Taxonomy Database (Chaumeil et al., 2019) (GTDB, v95), including marine representatives by using a genome sampling strategy reported previously (Martinez-Gutierrez and Aylward, 2021). In addition, we improved the representation of marine clades by subsampling genomes from the GORG database (Pachiadaki et al., 2019), which includes a wide range of genomes derived from single-cell sequencing, and adding several *Thermoarchaeota* genomes available on the JGI (Nordberg et al., 2014). We discarded genomes belonging to the DPANN superphylum due to the uncertainty of their placement within the archaea (Martinez-Gutierrez and Aylward, 2021). The list of genomes used is reported in Supplemental File 1.

We reconstructed a phylogenetic tree through the benchmarked MarkerFinder pipeline developed previously (Martinez-Gutierrez and Aylward, 2021), which resulted in an alignment of 27 ribosomal genes and three RNA polymerase genes (RNAP) (Martinez-Gutierrez and Aylward, 2021). The MarkerFinder pipeline consists of 1) the identification of ribosomal and RNAP genes using HMMER v3.2.1 with the reported model-specific cutoffs (Eddy, 2011; Sievers and Higgins, 2018), 2) alignment with ClustalOmega (Sievers and Higgins, 2018), and 3) concatenation of individual alignments. The resulting concatenated alignment was trimmed using trimAl (Capella-

Gutiérrez et al., 2009) with the option -gt 0.1. Phylogenetic tree inference was carried out with IQ-TREE v1.6.12 (Nguyen et al., 2015) with the options -wbt, -bb 1000 (Minh et al., 2013), -m LG+R10 (substitution model previously selected with the option -m MFP according to the Bayesian Information Criterion (Kalyaanamoorthy et al., 2017)), and --runs 5 to select the tree with the highest likelihood. The tree with the highest likelihood was manually inspected to discard the presence of topological inconsistencies and artifacts on iTOL (Letunic and Bork, 2019) (Fig .1). The raw phylogenetic tree is presented in Supplemental File 2. In a previous study, we assessed the effect of substitution model selection on the topology of a multi-domain phylogenetic tree (Martinez-Gutierrez and Aylward, 2021) however, we did not observe topological changes between the substitution models LG+C60 and LG+R10.

Assessment of Quality Tree

Due to the key importance of tree quality for the tree-dependent analysis performed in our study, we assessed the congruence of our prokaryotic ToL through the Tree Certainty metric (TC) (Martinez-Gutierrez and Aylward, 2021; Salichos et al., 2014), which has recently been shown to be a more accurate estimate for phylogenetic congruence that the traditional bootstrap. Our estimate based on 1,000 replicate trees (TC = 0.91) indicates high congruence in our phylogeny, indicating that the phylogenetic signal across our concatenated alignment of marker genes is consistent. We also evaluated whether the topology of our ToL agrees with a high-quality prokaryotic ToL reported previously (Martinez-Gutierrez and Aylward, 2021). In general, we observed consistency in the placement of all the phyla, as well as the bacterial superphyla (Terrabacteria and Gracilicutes) between both trees, except for the sisterhood of Actinobacteriota and Armatimonadota, which differs from the sisterhood of Actinobacteriota and Firmicutes in the

reference tree (Martinez-Gutierrez and Aylward, 2021). Despite this discrepancy, we do not expect that it will substantially impact the results of the current study because none of the marine clades are within this region of the tree.

Estimating the age of the crown node of bacterial and archaeal marine clades

To investigate the timing of the diversification of the marine planktonic clades, the phylogenetic tree obtained above was used to perform a molecular dating analysis of the crown nodes leading to the diversification of the main marine microbial clades. We focused our analysis on clades of bacteria and archaea that are overwhelmingly marine, such that the evolutionary history of that clade could be clearly traced back to an ancestral colonization of the ocean. Some clades, such as marine *Nitrospinae* and Actinobacteria, were not considered because they included several non-marine members, and it was unclear whether these lineages colonized the ocean multiple times independently. Our analysis was performed through Phylobayes v4.1c (Lartillot et al., 2009) with the program pb on four independent chains. For each chain, the input consisted in the phylogenetic tree, the amino acids alignment, the calibrations, and an autocorrelated relaxed log normal model (-ln) (Thorne et al., 1998) with the molecular evolution model CAT-Poisson+G4. Convergence was tested every 5000 cycles using the program tracecomp with a burn-in of 250 cycles and sampling every 2 cycles. After 100,000 cycles, our chains reached convergence in 8 out of 12 parameters (Supplemental File 3). To assess the uncertainty derived from the parameters that did not reach convergence, we estimated the divergence ages for each of our four chains using the last 1000 cycles and a range of 10 cycles to have a sample of 100 age estimates using the program readdiv (Supplemental File 4). Although some Bayesian parameters did not reach convergence after 100,000 cycles (Supplemental File 3 and 8), the estimated ages resulting from our four

independent chains were similar when compared to each other (Supplemental File 4). However, we observed an overall decrease in consistency between chains in the earliest clades (MGII, SAR324, Aerobic *Ca.* Marinimicrobia, and SAR202). This discrepancy is probably due to a decline in phylogenetic signal towards the root of the phylogenetic tree (Philippe et al., 2011).

Selection of priors and assessment of priors' impact on posterior distribution

To determine the impact of our priors on the age estimates of the calibrated nodes in our tree and assess the suitability of the ages used as priors for our analyses, we ran an independent MCMC chain without the amino acid alignment using the option -root on Phylobayes. Our prior-only analysis yielded a posterior age falling within the maximum and minimum priors used for the crown group of archaea and bacteria. For the internal calibrated nodes, we observed posterior estimates consistent with the priors used for each case except for aerobic ammonia oxidizing archaea (Supplemental File 5 and 6). Overall, this result suggests that the calibrations used as priors were adequate for our analyses.

Molecular dating analysis based on Penalized Likelihood and assessment of priors role on age estimates

We evaluated the reproducibility of our Bayesian divergence estimates by running an additional analysis based on Penalized Likelihood using TreePL (Smith and O'Meara, 2012) on 1000 replicate trees. Replicate trees were generated with the program bsBranchLenghts available on RAxML v8.2.12 (Stamatakis, 2014). For each replicate run, we initially used the option "prime" on TreePL to identify the optimization parameters and applied the parameters "through" to continue iterations until convergence in the parameters of each of the 1000 runs. Moreover, we

estimated the optional smoothing value for each replicate tree and ran cross-validation with the options “cv” and “randomcv”. The divergence times resulting from the 1000 bootstrap trees were used to assess the age variation for each marine microbial clade node (Supplemental File 4 and 8). Moreover, we used a Penalized Likelihood approach to assess the role of calibrations on age estimates by using two different sets of priors. The first set consisted in using the priors shown in Table 1 (Priors set 1) and the second set included the independent calibration of the bacterial and archaeal root and the crown node of oxygenic Cyanobacteria (Priors set 2).

Assessing the role of molecular dating strategy, molecular dating rate model, and calibrations on the diversification timing estimates of marine microbial clades

In order to evaluate the reproducibility of our Bayesian molecular dating analysis and to assess the reliability of the calibration points used (Figure 2), we applied multiple additional molecular dating analyses. Firstly, using the same calibrations, we applied a second independent approach based on Penalized Likelihood (PL) (Smith and O’Meara, 2012) on 1000 replicate bootstrap trees with fixed topology but varying branch lengths (See Methods). Estimates showed consistency in the age estimates between the two strategies tested except for the clades *Prochlorococcus* and *Synechococcus*, which showed a more recent diversification when using a Bayesian approach (Supplemental Fig 4). Secondly, we evaluated the role of model selection on our Bayesian posterior estimates by running two additional Bayesian analyses under the relaxed molecular clock models CIR (autocorrelated CIR process, (Lepage et al., 2007)) and UGAM (uncorrelated gamma multiplies, (Drummond et al., 2006)) available on Phylobayes. Overall, our estimates once again revealed broad consistency across models, with the exception that SAR11 and SAR86 had notably earlier divergence times with the CIR and UGAM models compared to the Log-normal model.

Previous research has shown that autocorrelated models outperform uncorrelated models when tested in different datasets (Lepage et al., 2007), which would suggest that the autocorrelated log-normal and CIR models provide the most robust estimates in our analysis. Indeed, for SAR86 the UGAM model provided an unusually early diversification date that is an outlier compared to all other estimates (Supplemental Fig 4).

Lastly, due to the potential limitations of using the oxygenation of the atmosphere (GOE) as a maximum prior for the strict aerobic metabolism of aerobic *Ca.* Marinimicrobia, Ammonia-Oxidizing Archaea, and Nitrate-Oxidizing Bacteria (Table 1), we performed an additional molecular dating analysis using a Penalized Likelihood approach in which these priors were excluded (Priors set 2; Supplemental File 8). Our analysis once again showed similar divergence times in all marine clades regardless of the priors used (Supplemental File 8), indicating that the use of these calibrations did not strongly shape our results. Importantly, the overall consistency in our age estimates using different molecular dating approaches, models, and priors does not alter our main conclusions regarding the emergence of marine microbial clades and the geochemical context in which they first diversified.

Comparing Bayesian diversification estimates with previous studies

Two estimated divergence times shown in our study disagree with previously published analyses. Firstly, a recent molecular dating estimate suggested that the transition of AOA-Archaea from terrestrial environments into marine reals occurred before the NOE (Ren et al., 2019; Smith and O'Meara, 2012) during a period known as the "boring million" characterized by low productivity and minimum oxygen concentrations in the atmosphere (0.1% the present levels) (Anbar and

Knoll, 2002; Hodgskiss et al., 2019; Holland, 2006; Reinhard and Planavsky, 2022). Our estimates point to a later diversification of this lineage during or after the NOE (678 Mya, 95% CI = 668-688 Mya) (Fig. 2), which is comparable with the age reported by another independent study (Yang et al., 2021). Secondly, another study reported the origin of the Picocyanobacterial clade *Prochlorococcus* to be 800 My, before the Snowball Earth period registered during the Cryogen[12]. However, our results agree with another independent study that points to a relatively late evolution of *Prochlorococcus* (Sánchez-Baracaldo, 2015).

Orthologous groups detection, stochastic mapping, and functional annotation

To investigate the genomic novelties associated with the diversification of the marine microbial lineages considered in our study, we identified enriched KEGG categories in the crown node of each clade. First, we predicted protein orthologous groups with ProteinOrtho v6 (Lechner et al., 2011) using the option "lastp" and protein files downloaded from the GTDB, GORG, and JGI databases. Furthermore, we performed a functional annotation using the KEGG database (Kanehisa, 2019; Kanehisa et al., 2021; Kanehisa and Goto, 2000) through HMMER3 with an e-value of 10^{-5} on all proteins. Proteins with multiple annotations were filtered to keep the best-scored annotation, and we predicted the function of each protein orthologous group by using the Majority Rule Principle. The presence/absence matrix resulting from the identification of orthologous groups was used together with the phylogenetic tree utilized for molecular dating to perform 100 replicate stochastic mapping analyses on each orthologous group with the make.simmap function implemented on Phytools (Bollback, 2006; Revell, 2012) with the model "all-rates-different" (ARD). To evaluate evidence of enrichment of KEGG categories, we created a null distribution for each protein cluster by simulating under the transition matrix estimated from

our stochastic mapping analysis using the function sim.history, but without conditioning on the presence/absence data at the tips (i.e. simulating a constant rate null distribution of transitions across the tree). Since some of the protein clusters show a low exchange rate (identified because one of the rows in the Q-matrix was equal to zero), we manually changed the exchange rate from zero to 0.00001. For each distribution, we estimated the number of genes gained for each KEGG category at the crown node of the marine clades. Clusters without a known annotation on the KEGG database were discarded. The resulting KEGG categories distributions for our stochastic mapping and null analyses were statistically compared using a one-tailed Wilcox test (α=0.01, N=100 for each distribution). KEGG categories showing statistically more gains in our stochastic mapping distribution were considered enriched (Supplemental File 7).

ACKNOWLEDGMENTS

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. This investigation was supported by grants from the Institute for Critical Technology and Applied Science and the National Science Foundation (IIBR-2141862), and a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A. We kindly thank members of the Aylward Lab for their insightful comments on an earlier version of this manuscript.

AUTHORS CONTRIBUTIONS

Conceived and designed this work: CAMG, UJC, and FOA. Wrote the manuscript: CAMG, UJC, and FOA.

SUPPLEMENTAL MATERIALS

Supplemental File 1. Genomes dataset used for the molecular dating of the main marine microbial clades.

Supplemental File 2. Raw maximum likelihood phylogenetic tree used for molecular dating and stochastic mapping analyses.

Supplemental File 3. Assessment of parameters convergence of four independent chains used for Bayesian molecular dating analyses. Relative difference <0.3 is shown in bold letters and denotes parameters that reached convergence after 100,000 cycles using a burn-in of 250 and sampling every two cycles.

Supplemental File 4. Comparison of the age distribution of marine microbial clades and its relationship with the main Earth oxygenation events using a Bayesian and a Penalized Likelihood approach for molecular dating. Ridges represent the age of 100 and 1000 replicate age estimates for each Bayesian independent chains and Penalized Likelihood analyses, respectively (see Methods).

Supplemental File 5. Estimated ages for calibrated nodes showing their suitability as priors for Bayesian molecular dating. Values resulted from running an independent chain on the temporal calibrations without sequence data (-root option on Phylobayes). Bars represent the standard error of the cycles tested.

Supplemental File 6. KOs gained at the crown node of each marine microbial clade. A KO was considered as gained when found in 51 out of 100 stochastic mapping replicates.

Supplemental File 7. Enriched KEGG categories at the crown node of each marine microbial clade. Clades were classified based on the diversification timing shown in Fig. 2. Enriched categories were identified by statistically comparing a stochastic mapping distribution with an all-rates-different vs a null distribution with a constant rate model without conditioning on the presence/absence data at the tips. Each dot represents one replicate (See Methods). X-axis represents the number of KOs gained at each crown node for each KEGG category. Stochastic mapping and null distributions were sorted for visualization purposes.

Supplemental File 8. Assessment of the role of molecular dating Bayesian model and calibrations on the diversification timing of marine microbial clades. Bayesian estimates represent the average of the last 1000 cycles sampled every 10 cycles of chain 1 (Supplemental File 9). TreePL analyses show 1000 age replicates using the priors shown in Table 1 (Priors set 1), and the independent root of Bacteria and Archaea and the minimum age of Cyanobacteria as priors (Priors set 2). Error bars represent the standard deviation of each distribution of replicate ages.

Supplemental File 9. Age estimates of marine microbial clades resulting from different Bayesian molecular dating models (Log-normal, CIR, and UGAM) and calibrations (TreePL priors set 1 and 2). Bayesian estimates represent the average of the last 1000 cycles sampled every 10 cycles of each of the four chains. TreePL analyses show 1000 age replicates using the priors shown in Table 1 (Priors set 1), and the independent root of Bacteria and Archaea and the minimum age of Cyanobacteria as priors (Priors set 2).

REFERENCES

Alcott LJ, Mills BJW, Poulton SW. 2019. Stepwise Earth oxygenation is an inherent property of global biogeochemical cycling. *Science* **366**:1333–1337.

Anbar AD, Duan Y, Lyons TW, Arnold GL, Kendall B, Creaser RA, Kaufman AJ, Gordon GW, Scott C, Garvin J, Buick R. 2007. A whiff of oxygen before the great oxidation event? *Science* **317**:1903–1906.

Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297**:1137–1142.

Bekker A, Holland HD, Wang P-L, Rumble D 3rd, Stein HJ, Hannah JL, Coetzee LL, Beukes NJ. 2004. Dating the rise of atmospheric oxygen. *Nature* **427**:117–120.

Bergman NM, (Tim) Lenton TM, Watson AJ, Dynamics B, Biogeochemistry. 2004. COPSE: A new model of biogeochemical cycling over Phanerozoic time.

Berner RA, Raiswell R. 1983. Burial of organic carbon and pyrite sulfur in sediments over phanerozoic time: a new theory. *Geochimica et Cosmochimica Acta*. doi:10.1016/0016-7037(83)90151-5

Biller SJ, Berube PM, Lindell D, Chisholm SW. 2015. Prochlorococcus: the structure and function of collective diversity. *Nat Rev Microbiol* **13**:13–27.

Bollback JP. 2006. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**:88.

Brasier MD, Lindsay JF. 1998. A billion years of environmental stability and the emergence of eukaryotes: new data from northern Australia. *Geology* **26**:555–558.

Brown MV, Ostrowski M, Grzymski JJ, Lauro FM. 2014. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. *Mar Genomics* **15**:17–28.

Burrows CJ. 2010. Surviving an Oxygen Atmosphere: DNA Damage and Repair. *ACS Symposium Series*. doi:10.1021/bk-2009-1025.ch008

Butterfield NJ. 2001. Paleobiology of the late Mesoproterozoic (ca. 1200 Ma) Hunting Formation, Somerset Island, arctic Canada. *Precambrian Research*. doi:10.1016/s0301-9268(01)00162-0

Butterfield NJ. 2000. *Bangiomorpha pubescens*n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology*. doi:10.1666/0094-8373(2000)026<0386:bpngns>2.0.co;2

Butterfield NJ, Knoll AH, Swett K. 2006. Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. Wiley-Blackwell.

Canfield DE, Poulton SW, Narbonne GM. 2007. Late-Neoproterozoic deep-ocean oxygenation

and the rise of animal life. *Science* **315**:92–95.
Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973.
Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**:1925–1927.
Clifford EL, Varela MM, De Corte D, Bode A, Ortiz V, Herndl GJ, Sintes E. 2019. Taurine Is a Major Carbon and Energy Source for Marine Prokaryotes in the North Atlantic Ocean off the Iberian Peninsula. *Microbial Ecology*. doi:10.1007/s00248-019-01320-y
Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. 2021. A rooted phylogeny resolves early bacterial evolution. *Science* **372**. doi:10.1126/science.abe0511
Crockford PW, Hayles JA, Bao H, Planavsky NJ, Bekker A, Fralick PW, Halverson GP, Bui TH, Peng Y, Wing BA. 2018. Triple oxygen isotope evidence for limited mid-Proterozoic primary productivity. *Nature* **559**:613–616.
Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**:90–93.
Dahl TW, Arens SKM. 2020. The impacts of land plant evolution on Earth's climate and oxygenation state – An interdisciplinary review. *Chemical Geology*. doi:10.1016/j.chemgeo.2020.119665
de la Torre JR, Christianson LM, Béjà O, Suzuki MT, Karl DM, Heidelberg J, DeLong EF. 2003. Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci U S A* **100**:12830–12835.
Dontsova K, Balogh-Brunstad Z, Le Roux G. 2020. Biogeochemical Cycles: Ecological Drivers and Environmental Impact. John Wiley & Sons.
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**:e88.
Ducklow HW, Doney SC. 2013. What Is the Metabolic State of the Oligotrophic Ocean? A Debate. *Annual Review of Marine Science*. doi:10.1146/annurev-marine-121211-172331
Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**:e1002195.
Falkowski PG, Barber RT, Smetacek V V. 1998. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**:200–207.
Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**:237–240.
Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WKW, Lomas MW, Veneziano D, Vera CS, Vrugt JA, Martiny AC. 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proc Natl Acad Sci U S A* **110**:9824–9829.
Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. *Nature*. doi:10.1038/nature04158
Hatzenpichler R. 2012. Diversity, physiology, and niche differentiation of ammonia-oxidizing archaea. *Appl Environ Microbiol* **78**:7501–7510.
Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR, Moisander PH, Paerl RW, Tripp HJ, Montoya JP, Moran MA, Zehr JP. 2009. In situ transcriptomic analysis of the globally important keystone N2-fixing taxon Crocosphaera watsonii. *ISME J* **3**:618–631.
Hodgskiss MSW, Crockford PW, Peng Y, Wing BA, Horner TJ. 2019. A productivity collapse to end Earth's Great Oxidation. *Proceedings of the National Academy of Sciences*.

doi:10.1073/pnas.1900325116
Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. 1998. A Neoproterozoic Snowball Earth. *Science*. doi:10.1126/science.281.5381.1342
Holland HD. 2006. The oxygenation of the atmosphere and oceans. *Philosophical Transactions of the Royal Society B: Biological Sciences*. doi:10.1098/rstb.2006.1838
Holland HD. 2002. Volcanic gases, black smokers, and the great oxidation event. *Geochimica et Cosmochimica Acta*. doi:10.1016/s0016-7037(02)00950-x
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589.
Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* **28**:1947–1951.
Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* **49**:D545–D551.
Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**:27–30.
Karl DM. 2002. Nutrient dynamics in the deep blue sea. *Trends Microbiol* **10**:410–418.
Khademian M, Imlay JA. 2021. How Microbes Evolved to Tolerate Oxygen. *Trends Microbiol* **29**:428–440.
Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006. Eukaryotic organisms in Proterozoic oceans. *Philos Trans R Soc Lond B Biol Sci* **361**:1023–1038.
Landry Z, Swan BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. 2017. SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. *MBio* **8**. doi:10.1128/mBio.00413-17
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**:2286–2288.
Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**:124.
Lenton TM, Dahl TW, Daines SJ, Mills BJW, Ozaki K, Saltzman MR, Porada P. 2016. Earliest land plants created modern levels of atmospheric oxygen. *Proc Natl Acad Sci U S A* **113**:9704–9709.
Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol* **24**:2669–2680.
Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*. doi:10.1093/nar/gkz239
Luo H, Csuros M, Hughes AL, Moran MA. 2013. Evolution of divergent life history strategies in marine alphaproteobacteria. *MBio* **4**. doi:10.1128/mBio.00373-13
Luo H, Moran MA. 2014. Evolutionary ecology of the marine Roseobacter clade. *Microbiol Mol Biol Rev* **78**:573–587.
Martinez-Gutierrez CA, Aylward FO. 2021. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol Biol Evol* **38**:5514–5527.
Masip L, Veeravalli K, Georgiou G. 2006. The many faces of glutathione in bacteria. *Antioxid Redox Signal* **8**:753–762.
Mason OU, Di Meo-Savoie CA, Van Nostrand JD, Zhou J, Fisk MR, Giovannoni SJ. 2009. Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. *ISME J* **3**:231–242.
McParland EL, Alexander H, Johnson WM. 2021. The Osmolyte Ties That Bind: Genomic

Insights Into Synthesis and Breakdown of Organic Osmolytes in Marine Microbes. *Frontiers in Marine Science*. doi:10.3389/fmars.2021.689306

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* **30**:1188–1195.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. 2004. High rates of N2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**:1027–1032.

Mühlenbruch M, Grossart H-P, Eigemann F, Voss M. 2018. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environ Microbiol* **20**:2671–2685.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**:268–274.

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* **42**:D26–31.

Och LM, Shields-Zhou GA. 2012. The Neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling. *Earth-Science Reviews*. doi:10.1016/j.earscirev.2011.09.004

Ossa Ossa F, Hofmann A, Spangenberg JE, Poulton SW, Stüeken EE, Schoenberg R, Eickmann B, Wille M, Butler M, Bekker A. 2019. Limited oxygen production in the Mesoarchean ocean. *Proc Natl Acad Sci U S A* **116**:6647–6652.

Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, Stepanauskas R. 2019. Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**:1623–1635.e11.

Pajares S, Varona-Cordero F, Hernández-Becerril DU. 2020. Spatial Distribution Patterns of Bacterioplankton in the Oxygen Minimum Zone of the Tropical Mexican Pacific. *Microb Ecol* **80**:519–536.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci U S A* **108**:13624–13629.

Partensky F, Garczarek L. 2010. Prochlorococcus: advantages and limits of minimalism. *Ann Rev Mar Sci* **2**:305–331.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9**:e1000602.

Planavsky NJ, Crowe SA, Fakhraee M, Beaty B, Reinhard CT, Mills BJW, Holstege C, Konhauser KO. 2021. Evolution of the structure and impact of Earth's biosphere. *Nature Reviews Earth & Environment*. doi:10.1038/s43017-020-00116-w

Planavsky NJ, Reinhard CT, Wang X, Thomson D, McGoldrick P, Rainbird RH, Johnson T, Fischer WW, Lyons TW. 2014. Earth history. Low mid-Proterozoic atmospheric oxygen levels and the delayed rise of animals. *Science* **346**:635–638.

Porter SM. 2004. The fossil record of early eukaryotic diversification. *The Paleontological Society Papers*. doi:10.1017/s1089332600002321

Reinhard CT, Planavsky NJ. 2022. The History of Ocean Oxygenation. *Ann Rev Mar Sci* **14**:331–353.

Ren M, Feng X, Huang Y, Wang H, Hu Z, Clingenpeel S, Swan BK, Fonseca MM, Posada D, Stepanauskas R, Hollibaugh JT, Foster PG, Woyke T, Luo H. 2019. Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J* **13**:2150–2161.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*. doi:10.1111/j.2041-210x.2011.00169.x

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* **31**:1261–1271.

Sánchez-Baracaldo P. 2015. Origin of marine planktonic cyanobacteria. *Sci Rep* **5**:17418.

Sánchez-Baracaldo P, Bianchini G, Di Cesare A, Callieri C, Chrismas NAM. 2019. Insights Into the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA and cpeBA). *Frontiers in Microbiology*. doi:10.3389/fmicb.2019.00045

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF, Hagemann M, Paulsen I, Partensky F. 2009. Ecological Genomics of Marine Picocyanobacteria. *Microbiology and Molecular Biology Reviews*. doi:10.1128/mmbr.00035-08

Scott C, Lyons TW, Bekker A, Shen Y, Poulton SW, Chu X, Anbar AD. 2008. Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452**:456–459.

Seymour JR, Amin SA, Raina J-B, Stocker R. 2017. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology*. doi:10.1038/nmicrobiol.2017.65

Shang H, Rothman DH, Fournier GP. 2022. Oxidative metabolisms catalyzed Earth's oxygenation. *Nat Commun* **13**:1328.

Sheik CS, Jain S, Dick GJ. 2014. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. *Environ Microbiol* **16**:304–317.

Shields-Zhou G, Och L. 2011. The case for a Neoproterozoic Oxygenation Event: Geochemical evidence and biological consequences. *GSA Today*. doi:10.1130/gsatg102a.1

Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **27**:135–145.

Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**:2689–2690.

Sperling EA, Wolock CJ, Morgan AS, Gill BC, Kunzmann M, Halverson GP, Macdonald FA, Knoll AH, Johnston DT. 2015. Statistical analysis of iron geochemical data suggests limited late Proterozoic oxygenation. *Nature*. doi:10.1038/nature14589

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Tang D, Shi X, Wang X, Jiang G. 2016. Extremely low oxygen concentration in mid-Proterozoic shallow seawaters. *Precambrian Research*. doi:10.1016/j.precamres.2016.02.005

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*. doi:10.1093/oxfordjournals.molbev.a025892

Thrash JC, Seitz KW, Baker BJ, Temperton B, Gillies LE, Rabalais NN, Henrissat B, Mason OU. 2017. Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone." *MBio* **8**. doi:10.1128/mBio.01017-17

Tostevin R, Mills BJW. 2020. Reconciling proxy records and models of Earth's oxygenation during the Neoproterozoic and Palaeozoic. *Interface Focus* **10**:20190137.

Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**:516–519.

Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ. 2012. Microbial oceanography of anoxic oxygen minimum zones. *Proc Natl Acad Sci U S A* **109**:15996–16003.

Valley JW, Cavosie AJ, Ushikubo T, Reinhard DA, Lawrence DF, Larson DJ, Clifton PH, Kelly TF, Wilde SA, Moser DE, Spicuzza MJ. 2014. Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. *Nature Geoscience*. doi:10.1038/ngeo2075

Vidal G, Moczydłowska-Vidal M. 1997. Biodiversity, speciation, and extinction trends of Proterozoic and Cambrian phytoplankton. *Paleobiology*. doi:10.1017/s0094837300016808

Vila-Costa M, Simó R, Harada H, Gasol JM, Slezak D, Kiene RP. 2006. Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science* **314**:652–654.

Walter M R, R Buick, JSR Dunlop. 1980. Stromatolites 3,400–3,500 Myr old from the North pole area, Western Australia. *Nature* **284**: 443-445.

Ward LM, Kirschvink JL, Fischer WW. 2016. Timescales of Oxygenation Following the Evolution of Oxygenic Photosynthesis. *Orig Life Evol Biosph* **46**:51–65.

Wei G-Y, Planavsky NJ, He T, Zhang F, Stockey RG, Cole DB, Lin Y-B, Ling H-F. 2021. Global marine redox evolution from the late Neoproterozoic to the early Paleozoic constrained by the integration of Mo and U isotope records. *Earth-Science Reviews*. doi:10.1016/j.earscirev.2021.103506

Wilson ST, Aylward FO, Ribalet F, Barone B, Casey JR, Connell PE, Eppley JM, Ferrón S, Fitzsimmons JN, Hayes CT, Romano AE, Turk-Kubo KA, Vislova A, Armbrust EV, Caron DA, Church MJ, Zehr JP, Karl DM, DeLong EF. 2017. Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nat Microbiol* **2**:17118.

Yang Y, Zhang C, Lenton TM, Yan X, Zhu M, Zhou M, Tao J, Phelps TJ, Cao Z. 2021. The Evolution Pathway of Ammonia-Oxidizing Archaea Shaped by Major Geological Events. *Mol Biol Evol* **38**:3637–3648.

Zehr JP, Kudela RM. 2011. Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann Rev Mar Sci* **3**:197–225.

Zhang H, Sun Y, Zeng Q, Crowe SA, Luo H. 2021. Snowball Earth, population bottleneck and evolution. *Proc Biol Sci* **288**:20211956.