# 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 colonized contemporary ocean environments remains unclear. Here, we reconstructed the geological dates 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 is the result of three main phases of colonization that coincide with major oxygenation events. The first phase took place after the initial oxygenation of the planet that occurred at the time of the Great Oxidation Event (2.4-2.2 Ga), after which several lineages that proliferate in oxygen minimum zones today first colonized marine niches. The second phase began around the time of the Neoproterozoic Oxidation Event (0.8-0.4 Ga) and included the diversification of the most abundant heterotrophic bacterial clades, consistent with the hypothesis that their diversification is linked to the emergence of large eukaryotic phytoplankton. The last phase encompasses prevalent cyanobacterial lineages and occurred after the Phanerozoic Oxidation Event (0.45-0.4 Ga), coinciding with the formation of the contemporary oligotrophic ocean. Our work clarifies the timing at which abundant lineages of bacteria and archaea colonized the ocean, links their adaptive radiations with key geological events, and

demonstrates that the redox state of the ocean throughout Earth's history has been the primary factor shaping the diversification of the most prevalent marine microbial clades.

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

## MAIN TEXT

The ocean plays a central role in the functioning and stability of Earth's biogeochemistry[1–3]. Due to their abundance, diversity, and physiological versatility, microbes mediate the vast majority of organic matter transformations that underpin higher trophic levels[1,2,4,5]. For example, marine microorganisms regulate a large fraction of the organic carbon pool[6], drive elemental cycling of nutrients like nitrogen[7], and participate in the ocean-atmosphere exchange of climatically important gasses[8]. 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[9]. Phylogenetic studies showed that these clades are broadly distributed across the Tree of Life (ToL) and encompass a wide range of phylogenetic breadths[9]. 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)[10,11], picocyanobacteria of the genera *Synechococcus* and *Prochlorococcus*[12–14], and marine alphaproteobacterial groups that included the SAR11 and Roseobacter clades[15]. Differences in the methodological frameworks used in these studies often hinder comparisons between lineages, however, and results for individual clades often conflict[10–13]. 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 address these challenges, 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 known for their abundance and major roles in marine biogeochemical cycles in the modern ocean (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[16] (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 superphyla in Bacteria and the basal placement of the *Thermatogales*[16,17] (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 taxonomic groups, 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 the age of the presence of liquid water as approximated through the dating of zircons[18], as well as the most ancient record of biogenic methane[18,19] 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)[20–22]. Similarly, we applied this reasoning for the calibration of the crown group of aerobic Ammonia Oxidizing Archaea, aerobic *Ca.* Marinimicrobia, and the Nitrite Oxidizing Bacteria, using their strict aerobic metabolism as evidence for a maximum of 2320 My. Our Bayesian estimates are consistent with 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 (2597 My; Fig. 2) is in agreement with geological evidence that points to their diversification happening before the GOE[23].

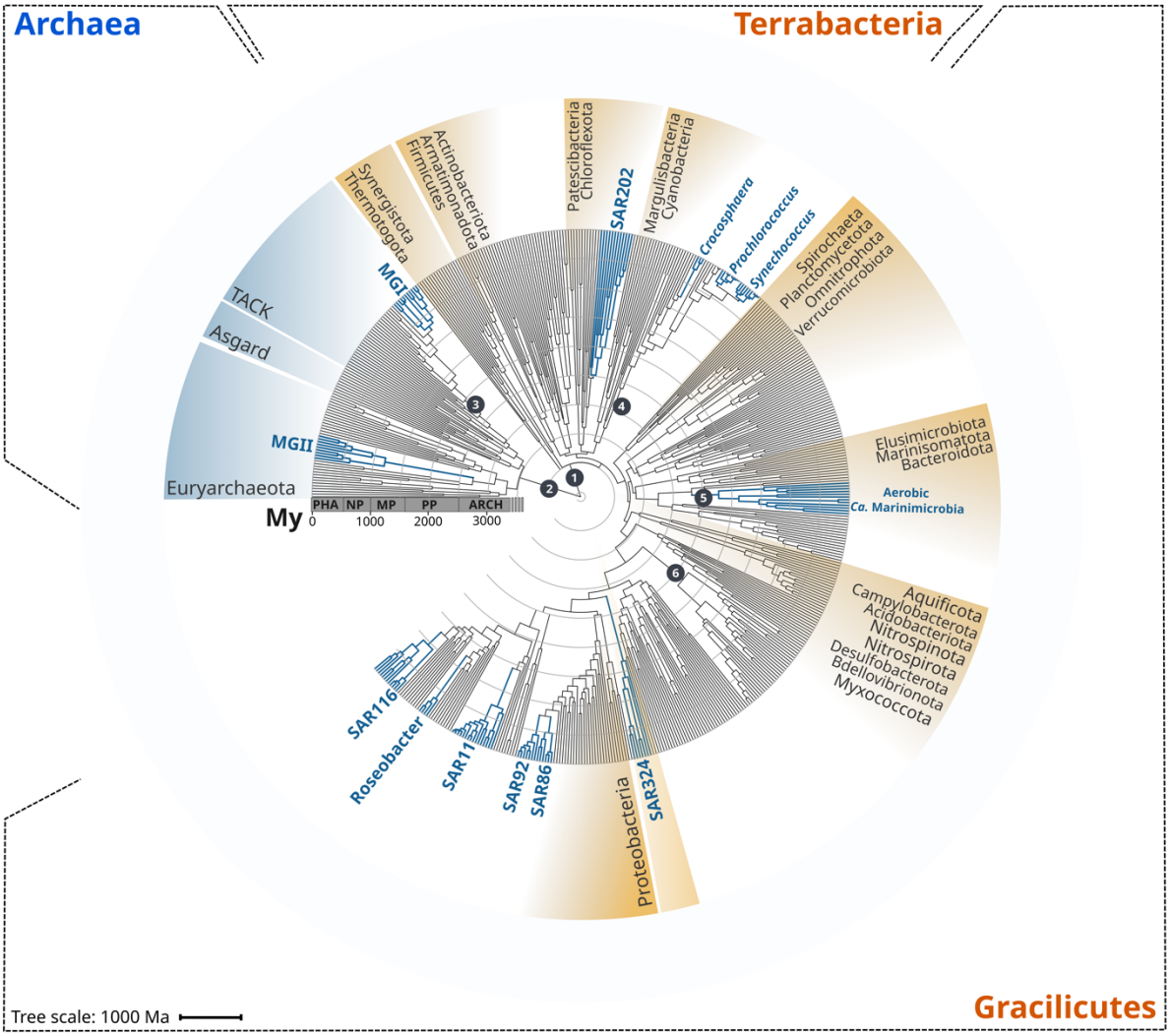



**Figure 1. Rooted inter-domain Tree of Life used for molecular dating analyses.** Maximum likelihood tree constructed in IQ-TREE v1.6.12 using the concatenation of 30 RNAP subunits and ribosomal protein sequences and the substitution model LG+R10. Blue labels represent the marine clades dated in our 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.**

| 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. | *17* |
| 1,2 | Bacteria-Archaea Root | 3460 | - | Identification of the most ancient traces of methane. | *18* |
| 3 | Aerobic *Nitrososphaerales* | - | 2320 | Strict aerobic metabolism. | *19* |
| 4 | Oxygenic Cyanobacteria | 2320 | - | Oxygenation of the atmosphere. The Great Oxidation Event has been associated with oxygenic Cyanobacteria. | *19* |
| 5 | Aerobic *Ca.* Marinimicrobia | - | 2320 | Strict aerobic metabolism. | *19* |
| 6 | Nitrite oxidizing bacteria | - | 2320 | Strict aerobic metabolism. | *19* |

The ages of marine bacterial and archaeal clades demonstrates that their diversification can be broadly divided into three phases that coincide with the major oxygenation events of the atmosphere and the ocean (Fig. 2). The first phase occurred near the time of the GOE, and included the clades SAR202, aerobic *Ca.* Marinimicrobia, SAR324, and the Marine Group II of the phylum Euryarchaeota (MGII). Within this first phase, the most ancient clade was the SAR202 (2479 My, 95% CI = 2465-2492 My), whose diversification took place near before the GOE (Fig. 2). The diversification of SAR202 before the rise of oxygen in the atmosphere suggests that this group emerged during an oxygen oasis proposed to have existed in pre-GOE Earth[24–26]. 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 by partially metabolizing 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[27,28]. After

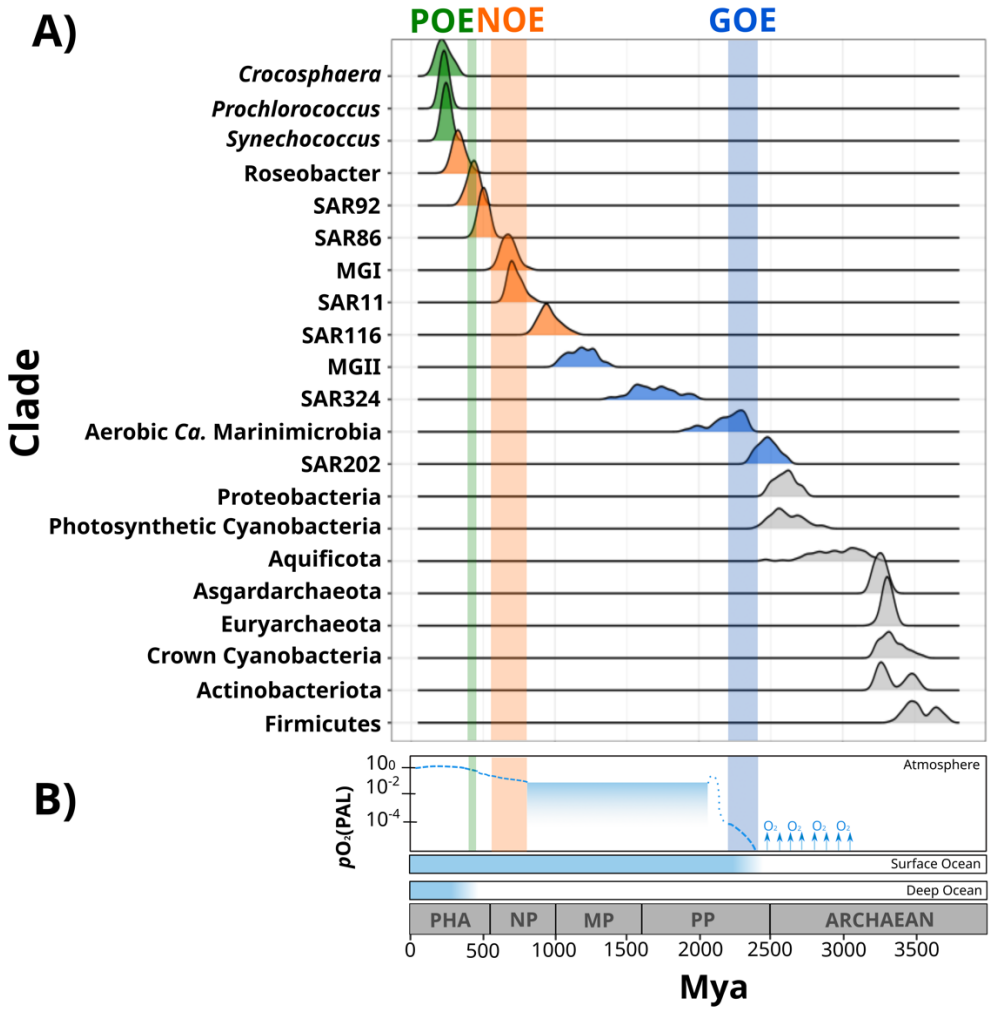
the GOE, we detected the diversification of aerobic *Ca.* Marinimicrobia (2196 My, 95% CI = 2173-2219 My), the SAR324 clade (1686 My, 95% CI = 1658-1715 My), and the MGII clade (1184 My, 95% CI = 1166-1202 My) (Fig. 2). Although these ancient clades may have first diversified in an oxic environment, the abrupt first increase of oxygen during the GOE was followed by a relatively rapid drop in ocean and atmosphere oxygen levels[25,29,30]. It is therefore likely that these clades diversified in the microaerophilic and variable oxygen conditions that prevailed during this period[20–22]. Indeed, the oxygen landscape in which these marine clades first diversified is consistent with their current physiology. There groups are capable of using oxygen as terminal electron acceptor however, they are prevalent in modern marine oxygen minimum (OMZs), where they use a wide range of alternative electron acceptors (e.g., nitrate and sulfate)[31–33]. The facultative aerobic or microaerophilic metabolism in these clades is therefore likely 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 phase, 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.

The second phase of diversification can be traced back to the time around the Neoproterozoic Oxygenation Event (POE) that occurred 800-540 My[30,34], and included the clades 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 relative late appearance of these heterotrophic lineages abundant in the open ocean today was likely 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 known as the "boring billion"[20,25,29,35–38]. The diversification of these clades may be indirectly associated with the tectonic activity and a Snowball event before the NOE[35,39,40], which increased the availability of terrestrial inorganic nutrients[35], and is also coincident with the widespread diversification of large eukaryotic algae during the Neoproterozoic[35,39,41–43]. 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 an increment in atmospheric oxygen concentrations[39,44]. 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[15,45]. Our results support this phenomenon and suggest that the interaction with marine Eukaryotes broadly influenced the diversification of several other heterotrophic lineages.

We also report the diversification of the chemolithoautotroph archaeal lineage MGI into the ocean during this second phase (678 My, 95% CI = 668-688 My) (Fig 2), which is comparable with the age reported by another independent study[10]. This is consistent with an increase in the oxygen concentrations of the ocean during this period[25], a necessary requisite for ammonia oxidation. Moreover, the widespread sulfidic conditions that have been proposed to have prevailed in the Mid-Proterozoic ocean likely limited the availability of redox-sensitive metals like Cupper, necessary for ammonia monooxygenases[35,46]. 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. Similar to what we observed in MGII and

169 SAR324, the Roseobacter clade shows a long branch leading to the crown node (Fig. 1), suggesting
170 that the diversification of this clade occurred earlier.
171




172

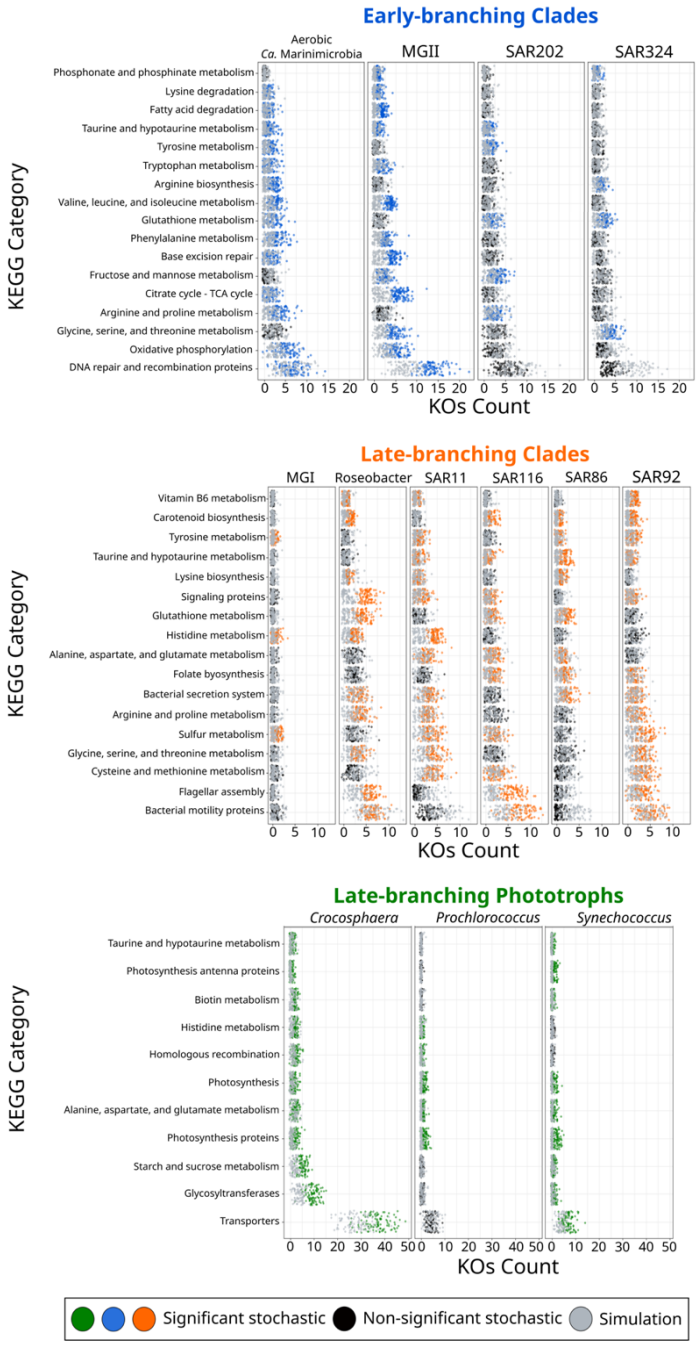
173 **Figure 2. Dates of the diversification of marine microbial clades and their relationship with**
174 **the redox history of Earth's atmosphere, surface ocean, and deep ocean.** A) Ridges represent
175 the distribution of 100 Bayesian dates estimated using a relaxed molecular clock and an
176 autocorrelated model (see Methods). Ridges of marine clades were colored based on their
177 diversification date: green, Late-branching Phototrophs; orange, Late-branching Clades; blue,
178 Early-branching Clades. The timing of the diversification of major bacterial and archaeal
179 superphyla is represented with gray ridges. B) Oxygenation events and redox changes across
180 Earth's history. Panel adapted from previous work[30]. Abbreviations: POE, Paleozoic Oxidation
181 Event; NOE, Neoproterozoic Oxidation Event; GOE, Great Oxidation Event; Pha, Paleozoic; NP,
182 Neoproterozoic; MP: Mesoproterozoic; PP: Paleoproterozoic.

The most recent and last phase of microbial diversification led to the appearance of late-branching phototrophs of the genera *Synechococcus* (243 My, 95% CI = 238-247 My), *Prochlorococcus* (230 My, 95% CI = 225-234 My), and the nitrogen-fixer *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*[13]. Picocyanobacteria and *Crocosphaera* are essential components of phytoplanktonic communities in the modern open ocean due to their large contribution to carbon and nitrogen fixation, respectively[47–49]. For example, the nitrogen fixation activities of *C. watsonii* in the open ocean today support the demands of nitrogen-starved microbial food webs found in oligotrophic waters[50]. The relatively late diversification of these lineages suggests that the oligotrophic open ocean is a relatively new 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[51]. 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[52–54]. Such widespread oxygenation was registered 430-390 My in an event referred to here as the Paleozoic Oxidation Event[55–58] (POE, Fig. 2). The increase of oxygen to present-day levels in the atmosphere and the ocean was the result of an increment of the burial of organic carbon in sedimentary rocks due to the diversification of the earliest land plants[25,57,59], which has been also associated with increased phosphorus weathering rates[57,60], global impacts on the global element cycles[61], and an increase in overall ocean productivity[59]. 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 possible once modern oxygen concentrations and biogeochemical dynamics were reached[25,51].

In order 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. Due to the timing of diversification events during the phases described previously, we classified the marine clades as Early-branching Clades, Late-branching Clades, and Late-branching Phototrophs based on their diversification timing (Fig. 2). To identify the enrichment of gene functions at the base 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 overall, 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[62–64]. Moreover, the EBC were enriched in proteins involved in ancient aerobic pathways, like oxidative phosphorylation and TCA cycle (Fig. 3), as well as genes implicated in the degradation of fatty acids under aerobic conditions, for example the enzyme alkane 1-monooxygenase in MGII (Supplemental File 6). We also detected genes for the adaptation to marine environments, for instance genes for the anabolism of taurine (e.g., cysteine dioxygenase in MGII, Supplemental File 6), an osmoprotectant commonly found in marine

228 bacteria[65]. Our findings suggest that the diversification of EBC in the ocean was linked to the
229 evolution of aerobic metabolism, the acquisition of metabolic capabilities to exploit the newly
230 created niches that followed the increase of oxygen, and the expansion of genes involved in the
231 tolerance to oxidative and salinity stress.
232
233




234

235 **Figure 3. KEGG categories enriched at the crown node of each marine microbial clade.**
236 Clades were classified based on their diversification timing shown in Fig. 2. Enriched categories
237 were identified by statistically comparing a stochastic mapping distribution with an all-rates-
238 different model vs a null distribution with a constant rate model without conditioning on the
239 presence/absence data at the tips. Each dot represents one replicate (See Methods). X-axis
240 represents the number of KOs gained at each crown node for each KEGG category. Stochastic
241 mapping and null distributions were sorted for visualization purposes. The complete list of
242 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[66], 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[67,68]. Although the phycosphere can also be found in prokaryotic phytoplankton[67], 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, as suggested by our Bayesian estimates (Fig. 2 and 4). The potential diversification of heterotrophic LBC due to their ecological proximity 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 interactions[67,69]. 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 substrate[70], instead of producing it as 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[71]. The enrichment of

genes potentially involved in bacteria-phytoplankton interactions suggest that the diversification of marine clades during the NOE was intimately linked to the establishment of ecological relationships with eukaryotes to exchange nutrients.

Late-branching phototrophs that diversified around the time of the POE (LBP; Fig. 2), showed the 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[50,72], which requires a rapid and efficient use of the scarce nutrients available. We also identified genes for osmotic pressure tolerance, for example a Ca-activated chloride channel homolog, a magnesium exporter, and a fluoride exporter (Supplemental File 6). This finding suggests that *Crocosphaera* might have diversified from a non-marine group into the ocean. 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). The absence of salt-tolerance related genes suggests that the ancestor of these Picocyanobacterial clades inhabited a low-nutrient marine habitat. 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[70]. *Prochlorococcus* exhibits enrichment in fewer categories than the rest of phototrophic clades diversifying during the same period, consistent with the streamlined nature of

289 genomes from this lineage[73]. The genes acquired by this lineage are involved in photosynthesis,
290 which supports previous findings that the diversification of this clade was accompanied by changes
291 in the photosynthetic apparatus compared with *Synechococcus*, its sister group[74]. Overall, the
292 diversification of LBP was marked by the capacity to thrive in the oligotrophic ocean by exploiting
293 organic and inorganic nutrients and modifying the photosynthetic apparatus as observed in
294 *Crocosphaera* and *Synechococcus*, and *Prochlorococcus,* respectively.
295
296 The contemporary ocean is dominated by abundant clades of bacteria and archaea that drive global
297 biogeochemical cycles and play a central role in shaping the redox state of the planet. Despite their
298 importance, the timing, and the geological landscape at which these clades colonized the ocean
299 have remained unclear due to a combination of the inherent difficulties of studying biological
300 events that occurred in deep time and the lack of a fossil record for microbial life. Here we present
301 a comprehensive timeline for the colonization of the ocean by abundant marine clades and reveal
302 that major oxygenation events in Earth's history played critical roles in creating new niches for
303 microbial diversification. These colonization events subsequently led to the establishment of the
304 biogeochemical cycles that govern the environmental stability of our planet today. Our study
305 allowed us to reconstruct a framework that links major geological and biological events to the
306 emergence of microbial lineages that dominate the contemporary ocean (Fig. 4). This provides key
307 foundational knowledge for understanding ongoing anthropogenic changes in the ocean.
308 Importantly, climate change is predicted to lead to an expansion of both oxygen minimum zones,
309 which represent refugia dating back to the mid-Proterozoic ocean, and oligotrophic surface waters,
310 which represent ecosystems that have emerged relatively recently in the Phanerozoic. Thus, the
311 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 critical context for understanding the impact of drastic climatic changes on the marine biome more broadly and will help to clarify how continuing ecological shifts will impact marine biogeochemical cycles.

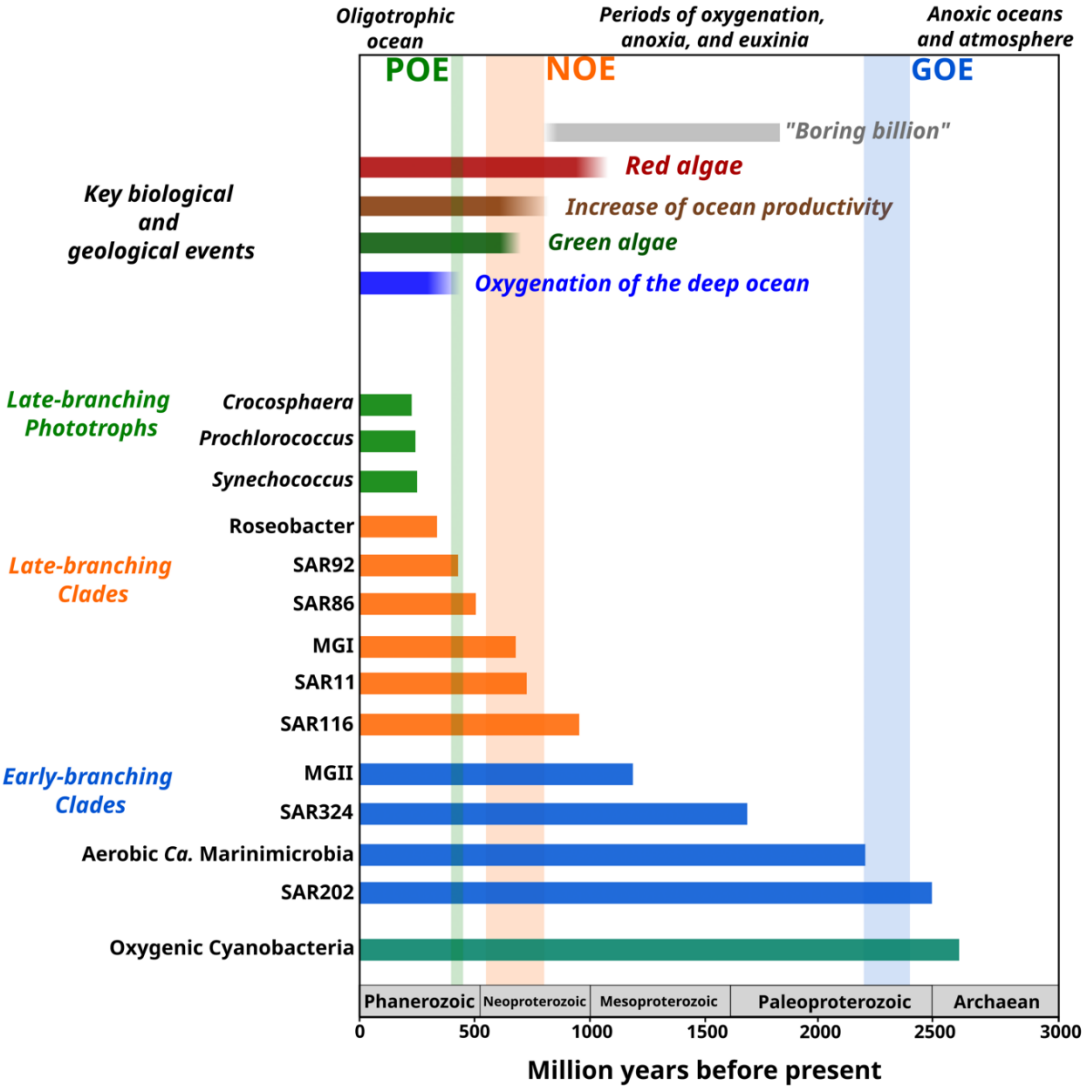



**Figure 4. Link between the timing of the diversification of the main marine microbial clades and major geological and biological events.** The timing of the geological and biological events potentially involved in the diversification of marine clades is based on previously published data: "Boring billion"[96], red algae fossils[97], increased of ocean productivity[34], green algae fossils[98], oxygenation of the deep ocean[57]. 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[30].

## MATERIAL AND METHODS

*Genomes sampling and phylogenetic reconstruction*

In order 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 subsampled from the Genome Taxonomy Database[75] (GTDB, v95), including marine representatives. In addition, we improved the representation of marine genomes by subsampling genomes from the GORG database[76], which includes a wide range of genomes derived from single-cell sequencing, and adding several *Thermoarchaeota* genomes available on the JGI[77]. We discarded genomes belonging to the DPANN superphylum due to the uncertainty of their placement within the archaea[16]. The list of genomes used is reported in Supplemental File 1.

We reconstructed a phylogenetic tree through the MarkerFinder pipeline developed previously[16], which resulted in an alignment of 27 ribosomal genes and three RNA polymerase genes (RNAP)[16]. The MarkerFinder pipeline consists of 1) the identification of ribosomal and RNAP genes using HMMER v 3.2.1 with the reported model-specific cutoffs[78], 2) alignment with ClustalOmega[78,79], and 3) concatenation of individual alignments. The resulting concatenated alignment was trimmed using trimAl[80] with the option -gt 0.1. Phylogenetic tree inference was carried out with IQ-TREE v1.6.12[81] with the options -wbt, -bb 1000[82], -m LG+R10 (substitution model previously selected with the option -m MFP according to the Bayesian Information Criterion[83]), 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[84] (Fig .1). The raw phylogenetic tree is presented in Supplemental File 2.

*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)[16,85], 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 is in agreement with a high-quality prokaryotic ToL reported previously[16]. In general, we observed consistency in the placement of the 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[16]. 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*

In order to investigate the timing of the diversification of marine planktonic clades of interest, the phylogenetic tree obtained previously was used to perform a molecular dating analysis of the crown nodes leading to the diversification of the main marine microbial clades. Our analysis was performed through Phylobayes v4.1c[86] 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)[87] with the default molecular evolution model. 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). In order 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). We report the confidence intervals and the median age of 100 replicates of chain three throughout the manuscript.

To determine the impact of our priors (Fig. 1 and Table 1) 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.

To evaluate the reproducibility of our Bayesian molecular dating analysis, we applied a second independent approach based on Penalized Likelihood (PL) using the TreePL[88] program on 1000 replicate bootstrap trees that had fixed topology but varying branch lengths. Replicate trees were generated with the program bsBranchLenghts available on RAxML v8.2.12[89]. For each replicate run, we initially used the option "prime" 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"[88]. The divergence times resulting from the

1000 bootstrap trees were used to assess the age variation for each node of interest (Supplemental File 4).

*Comparison among PhyloBayes chains and between TreePL age estimates*

Although some Bayesian parameters did not reach convergence after 100,000 cycles (Supplemental File 3), the estimated ages resulting from our four independent chains were similar when compared to each other (Supplemental File 4). Moreover, our Bayesian and Penalized likelihood approaches showed similar divergence times, strengthening the conclusions of our study. We only observed notable discrepancies between these approaches in Photosynthetic Cyanobacteria (PL showing more recent divergence during the GOE), and the marine Picocyanobacteria *Synechococcus* and *Prochlorococcus* (PL showing more ancient divergence during the POE). Despite these discrepancies, the differences observed between both approaches do not alter our main conclusions regarding the phases during which these major marine clades 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[11] during a period known as the "boring million" characterized by low productivity and minimum oxygen concentrations in the atmosphere (0.1% the present levels)[20,25,29,35]. 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[10]. 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*[13].

*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[90] 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[91–93] 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[94,95] 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

439 of the marine clades. Clusters without a known annotation on the KEGG database were discarded.
440 The resulting KEGG categories distributions for our stochastic mapping and null analyses were
441 statistically compared using a one-tailed Wilcox test ($\alpha$=0.01, N= 100 for each distribution).
442 KEGG categories showing statistically more gains in our stochastic mapping distribution were
443 considered enriched (Supplemental File 7).
444

445 **ACKNOWLEDGMENTS**

446 We acknowledge the use of the Virginia Tech Advanced Research Computing Center for
447 bioinformatic analyses performed in this study. This investigation was supported by grants from
448 the Institute for Critical Technology and Applied Science and the National Science Foundation
449 (IIBR-2141862), and a Simons Early Career Award in Marine Microbial Ecology and Evolution
450 to F.O.A. We kindly thank members of the Aylward Lab for their insightful comments on an earlier
451 version of this manuscript.
452

453 **AUTHORS CONTRIBUTIONS**

454 Conceived and designed this work: CAMG, UJC, and FOA. Wrote the manuscript: CAMG, UJC,
455 and FOA.
456

457 **MATERIALS AND CORRESPONDACE**

458 Frank O. Aylward; faylward@vt.edu

459 Carolina A. Martinez-Gutierrez; cmartinez@vt.edu
460
461

# 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. KEGG categories enriched 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.

## REFERENCES

1. Falkowski, P. G., Barber, R. T. & Smetacek, V., V. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200–207 (1998).
2. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
3. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
4. Mason, O. U. *et al.* Prokaryotic diversity, distribution, and insights into their role in biogeochemical cycling in marine basalts. *ISME J.* **3**, 231–242 (2009).
5. Brown, M. V., Ostrowski, M., Grzymski, J. J. & Lauro, F. M. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades. *Mar. Genomics* **15**, 17–28 (2014).
6. Ducklow, H. W. & Doney, S. C. What Is the Metabolic State of the Oligotrophic Ocean? A Debate. *Annual Review of Marine Science* vol. 5 525–533 Preprint at https://doi.org/10.1146/annurev-marine-121211-172331 (2013).
7. Zehr, J. P. & Kudela, R. M. Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann. Rev. Mar. Sci.* **3**, 197–225 (2011).
8. Vila-Costa, M. *et al.* Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science* **314**, 652–654 (2006).
9. Giovannoni, S. J. & Stingl, U. Molecular diversity and ecology of microbial plankton. *Nature* vol. 437 343–348 Preprint at https://doi.org/10.1038/nature04158 (2005).
10. Yang, Y. *et al.* The Evolution Pathway of Ammonia-Oxidizing Archaea Shaped by Major Geological Events. *Mol. Biol. Evol.* **38**, 3637–3648 (2021).
11. Ren, M. *et al.* Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J.* **13**, 2150–2161 (2019).
12. Zhang, H., Sun, Y., Zeng, Q., Crowe, S. A. & Luo, H. Snowball Earth, population bottleneck and *Prochlorococcus* evolution. *Proceedings of the Royal Society B: Biological Sciences* vol. 288 Preprint at https://doi.org/10.1098/rspb.2021.1956 (2021).
13. Sánchez-Baracaldo, P. Origin of marine planktonic cyanobacteria. *Sci. Rep.* **5**, 17418 (2015).
14. Sánchez-Baracaldo, P., Bianchini, G., Di Cesare, A., Callieri, C. & Chrismas, N. A. M. Insights Into the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA and

cpeBA). *Frontiers in Microbiology* vol. 10 Preprint at https://doi.org/10.3389/fmicb.2019.00045 (2019).

15. Luo, H., Csuros, M., Hughes, A. L. & Moran, M. A. Evolution of divergent life history strategies in marine alphaproteobacteria. *MBio* **4**, (2013).
16. Martinez-Gutierrez, C. A. & Aylward, F. O. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* **38**, 5514–5527 (2021).
17. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, (2021).
18. Valley, J. W. *et al.* Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. *Nature Geoscience* vol. 7 219–223 Preprint at https://doi.org/10.1038/ngeo2075 (2014).
19. Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S. & Isozaki, Y. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**, 516–519 (2006).
20. Holland, H. D. The oxygenation of the atmosphere and oceans. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 361 903–915 Preprint at https://doi.org/10.1098/rstb.2006.1838 (2006).
21. Holland, H. D. Volcanic gases, black smokers, and the great oxidation event. *Geochimica et Cosmochimica Acta* vol. 66 3811–3826 Preprint at https://doi.org/10.1016/s0016-7037(02)00950-x (2002).
22. Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* vol. 427 117–120 Preprint at https://doi.org/10.1038/nature02260 (2004).
23. Ward, L. M., Kirschvink, J. L. & Fischer, W. W. Timescales of Oxygenation Following the Evolution of Oxygenic Photosynthesis. *Orig. Life Evol. Biosph.* **46**, 51–65 (2016).
24. Ossa Ossa, F. *et al.* Limited oxygen production in the Mesoarchean ocean. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6647–6652 (2019).
25. Reinhard, C. T. & Planavsky, N. J. The History of Ocean Oxygenation. *Ann. Rev. Mar. Sci.* **14**, 331–353 (2022).
26. Anbar, A. D. *et al.* A whiff of oxygen before the great oxidation event? *Science* **317**, 1903–1906 (2007).
27. Landry, Z., Swan, B. K., Herndl, G. J., Stepanauskas, R. & Giovannoni, S. J. SAR202

Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. *MBio* **8**, (2017).

28. Shang, H., Rothman, D. H. & Fournier, G. P. Oxidative metabolisms catalyzed Earth's oxygenation. *Nature Communications* vol. 13 Preprint at https://doi.org/10.1038/s41467-022-28996-0 (2022).
29. Hodgskiss, M. S. W., Crockford, P. W., Peng, Y., Wing, B. A. & Horner, T. J. A productivity collapse to end Earth's Great Oxidation. *Proceedings of the National Academy of Sciences* vol. 116 17207–17212 Preprint at https://doi.org/10.1073/pnas.1900325116 (2019).
30. Alcott, L. J., Mills, B. J. W. & Poulton, S. W. Stepwise Earth oxygenation is an inherent property of global biogeochemical cycling. *Science* **366**, 1333–1337 (2019).
31. Thrash, J. C. *et al.* Metabolic roles of uncultivated bacterioplankton lineages in the northern Gulf of Mexico 'Dead Zone'. Preprint at https://doi.org/10.1101/095471.
32. Pajares, S., Varona-Cordero, F. & Hernández-Becerril, D. U. Spatial Distribution Patterns of Bacterioplankton in the Oxygen Minimum Zone of the Tropical Mexican Pacific. *Microb. Ecol.* **80**, 519–536 (2020).
33. Sheik, C. S., Jain, S. & Dick, G. J. Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. *Environmental Microbiology* vol. 16 304–317 Preprint at https://doi.org/10.1111/1462-2920.12165 (2014).
34. Och, L. M. & Shields-Zhou, G. A. The Neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling. *Earth-Science Reviews* vol. 110 26–57 Preprint at https://doi.org/10.1016/j.earscirev.2011.09.004 (2012).
35. Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297**, 1137–1142 (2002).
36. Tang, D., Shi, X., Wang, X. & Jiang, G. Extremely low oxygen concentration in mid-Proterozoic shallow seawaters. *Precambrian Research* vol. 276 145–157 Preprint at https://doi.org/10.1016/j.precamres.2016.02.005 (2016).
37. Crockford, P. W. *et al.* Triple oxygen isotope evidence for limited mid-Proterozoic primary productivity. *Nature* vol. 559 613–616 Preprint at https://doi.org/10.1038/s41586-018-0349-y (2018).
38. Planavsky, N. J. *et al.* Earth history. Low mid-Proterozoic atmospheric oxygen levels and the delayed rise of animals. *Science* **346**, 635–638 (2014).

595 39. Shields-Zhou, G. & Och, L. The case for a Neoproterozoic Oxygenation Event: Geochemical
596 evidence and biological consequences. *GSA Today* vol. 21 4–11 Preprint at
597 https://doi.org/10.1130/gsatg102a.1 (2011).
598 40. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic Snowball
599 Earth. *Science* vol. 281 1342–1346 Preprint at https://doi.org/10.1126/science.281.5381.1342
600 (1998).
601 41. Butterfield, N. J. Paleobiology of the late Mesoproterozoic (ca. 1200 Ma) Hunting Formation,
602 Somerset Island, arctic Canada. *Precambrian Research* vol. 111 235–256 Preprint at
603 https://doi.org/10.1016/s0301-9268(01)00162-0 (2001).
604 42. Vidal, G. & Moczydłowska-Vidal, M. Biodiversity, speciation, and extinction trends of
605 Proterozoic and Cambrian phytoplankton. *Paleobiology* vol. 23 230–246 Preprint at
606 https://doi.org/10.1017/s0094837300016808 (1997).
607 43. Porter, S. M. The fossil record of early eukaryotic diversification. *The Paleontological*
608 *Society Papers* vol. 10 35–50 Preprint at https://doi.org/10.1017/s1089332600002321 (2004).
609 44. Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. Eukaryotic organisms in Proterozoic
610 oceans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1023–1038 (2006).
611 45. Luo, H. & Moran, M. A. Evolutionary ecology of the marine Roseobacter clade. *Microbiol.*
612 *Mol. Biol. Rev.* **78**, 573–587 (2014).
613 46. Hatzenpichler, R. Diversity, physiology, and niche differentiation of ammonia-oxidizing
614 archaea. *Appl. Environ. Microbiol.* **78**, 7501–7510 (2012).
615 47. Scanlan, D. J. *et al.* Ecological Genomics of Marine Picocyanobacteria. *Microbiology and*
616 *Molecular Biology Reviews* vol. 73 249–299 Preprint at https://doi.org/10.1128/mmbr.00035-
617 08 (2009).
618 48. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria
619 *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences* vol.
620 110 9824–9829 Preprint at https://doi.org/10.1073/pnas.1307701110 (2013).
621 49. Montoya, J. P. *et al.* High rates of N2 fixation by unicellular diazotrophs in the oligotrophic
622 Pacific Ocean. *Nature* **430**, 1027–1032 (2004).
623 50. Hewson, I. *et al.* In situ transcriptomic analysis of the globally important keystone N2-fixing
624 taxon Crocosphaera watsonii. *ISME J.* **3**, 618–631 (2009).
625 51. Karl, D. M. Nutrient dynamics in the deep blue sea. *Trends Microbiol.* **10**, 410–418 (2002).

626 52. Scott, C. *et al.* Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452**, 456–
627 459 (2008).
628 53. Canfield, D. E., Poulton, S. W. & Narbonne, G. M. Late-Neoproterozoic Deep-Ocean
629 Oxygenation and the Rise of Animal Life. *Science* vol. 315 92–95 Preprint at
630 https://doi.org/10.1126/science.1135013 (2007).
631 54. Wei, G.-Y. *et al.* Global marine redox evolution from the late Neoproterozoic to the early
632 Paleozoic constrained by the integration of Mo and U isotope records. *Earth-Science Reviews*
633 vol. 214 103506 Preprint at https://doi.org/10.1016/j.earscirev.2021.103506 (2021).
634 55. Sperling, E. A. *et al.* Statistical analysis of iron geochemical data suggests limited late
635 Proterozoic oxygenation. *Nature* vol. 523 451–454 Preprint at
636 https://doi.org/10.1038/nature14589 (2015).
637 56. Berner, R. A. & Raiswell, R. Burial of organic carbon and pyrite sulfur in sediments over
638 phanerozoic time: a new theory. *Geochimica et Cosmochimica Acta* vol. 47 855–862 Preprint
639 at https://doi.org/10.1016/0016-7037(83)90151-5 (1983).
640 57. Lenton, T. M. *et al.* Earliest land plants created modern levels of atmospheric oxygen. *Proc.*
641 *Natl. Acad. Sci. U. S. A.* **113**, 9704–9709 (2016).
642 58. Tostevin, R. & Mills, B. J. W. Reconciling proxy records and models of Earth's oxygenation
643 during the Neoproterozoic and Palaeozoic. *Interface Focus* vol. 10 20190137 Preprint at
644 https://doi.org/10.1098/rsfs.2019.0137 (2020).
645 59. Planavsky, N. J. *et al.* Evolution of the structure and impact of Earth's biosphere. *Nature*
646 *Reviews Earth & Environment* vol. 2 123–139 Preprint at https://doi.org/10.1038/s43017-
647 020-00116-w (2021).
648 60. Bergman, N. M., (Tim) Lenton, T. M., Watson, A. J., Dynamics, B. & Biogeochemistry.
649 *COPSE: A new model of biogeochemical cycling over Phanerozoic time*. (2004).
650 61. Dahl, T. W. & Arens, S. K. M. The impacts of land plant evolution on Earth's climate and
651 oxygenation state – An interdisciplinary review. *Chemical Geology* vol. 547 119665 Preprint
652 at https://doi.org/10.1016/j.chemgeo.2020.119665 (2020).
653 62. Burrows, C. J. Surviving an Oxygen Atmosphere: DNA Damage and Repair. *ACS Symp. Ser.*
654 *Am. Chem. Soc.* **2009**, 147–156 (2009).
655 63. Masip, L., Veeravalli, K. & Georgiou, G. The many faces of glutathione in bacteria. *Antioxid.*
656 *Redox Signal.* **8**, 753–762 (2006).

64. Khademian, M. & Imlay, J. A. How Microbes Evolved to Tolerate Oxygen. *Trends Microbiol.* **29**, 428–440 (2021).
65. McParland, E. L., Alexander, H. & Johnson, W. M. The Osmolyte Ties That Bind: Genomic Insights Into Synthesis and Breakdown of Organic Osmolytes in Marine Microbes. *Frontiers in Marine Science* vol. 8 Preprint at https://doi.org/10.3389/fmars.2021.689306 (2021).
66. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13624–13629 (2011).
67. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology* vol. 2 Preprint at https://doi.org/10.1038/nmicrobiol.2017.65 (2017).
68. Mühlenbruch, M., Grossart, H.-P., Eigemann, F. & Voss, M. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environ. Microbiol.* **20**, 2671–2685 (2018).
69. Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J. & Smith, A. G. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**, 90–93 (2005).
70. Clifford, E. L. *et al.* Taurine Is a Major Carbon and Energy Source for Marine Prokaryotes in the North Atlantic Ocean off the Iberian Peninsula. *Microb. Ecol.* **78**, 299–312 (2019).
71. de la Torre, J. R. *et al.* Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12830–12835 (2003).
72. Wilson, S. T. *et al.* Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium Crocosphaera. *Nat Microbiol* **2**, 17118 (2017).
73. Partensky, F. & Garczarek, L. Prochlorococcus: advantages and limits of minimalism. *Ann. Rev. Mar. Sci.* **2**, 305–331 (2010).
74. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13–27 (2015).
75. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz848.
76. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-

Cell Genomics. *Cell* **179**, 1623–1635.e11 (2019).

77. Nordberg, H. *et al.* The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42**, D26–31 (2014).
78. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
79. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
80. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
81. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
82. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
83. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
84. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
85. Salichos, L., Stamatakis, A. & Rokas, A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol. Biol. Evol.* **31**, 1261–1271 (2014).
86. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
87. Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647–1657 (1998).
88. Smith, S. A. & O'Meara, B. C. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
89. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
90. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).

719 91. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
720 *Res.* **28**, 27–30 (2000).
721 92. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein*
722 *Sci.* **28**, 1947–1951 (2019).
723 93. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG:
724 integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
725 94. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things).
726 *Methods in Ecology and Evolution* vol. 3 217–223 Preprint at https://doi.org/10.1111/j.2041-
727 210x.2011.00169.x (2012).
728 95. Bollback, J. P. SIMMAP: stochastic character mapping of discrete traits on phylogenies.
729 *BMC Bioinformatics* **7**, 88 (2006).
730 96. Brasier, M. D. & Lindsay, J. F. A billion years of environmental stability and the emergence
731 of eukaryotes: new data from northern Australia. *Geology* **26**, 555–558 (1998).
732 97. Butterfield, N. J. *Bangiomorpha pubescens*n. gen., n. sp.: implications for the evolution of
733 sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes.
734 *Paleobiology* vol. 26 386–404 Preprint at https://doi.org/10.1666/0094-
735 8373(2000)026<0386:bpngns>2.0.co;2 (2000).
736 98. Moczydłowska, M. Paleobiology of the neoproterozoic svanbergfjellet formation,
737 spitsbergen. *Palaeogeography, Palaeoclimatology, Palaeoecology* vol. 122 247–248 Preprint
738 at https://doi.org/10.1016/0031-0182(96)85042-5 (1996).

739