ARTICLES
https://doi.org/10.1038/s41559-020-01386-9

nature ecology & evolution

The evolution of oxygen-utilizing enzymes suggests early biosphere oxygenation

Jagoda Jabłońska and Dan S. Tawfik ✉

Production of molecular oxygen was a turning point in the Earth's history. The geological record indicates the Great Oxidation Event, which marked a permanent transition to an oxidizing atmosphere around 2.4 Ga. However, the degree to which oxygen was available to life before oxygenation of the atmosphere remains unknown. Here, phylogenetic analysis of all known oxygen-utilizing and -producing enzymes (O_2-enzymes) indicates that oxygen became widely available to living organisms well before the Great Oxidation Event. About 60% of the O_2-enzyme families whose birth can be dated appear to have emerged at the separation of terrestrial and marine bacteria (22 families, compared to two families assigned to the last universal common ancestor). This node, dubbed the last universal oxygen ancestor, coincides with a burst of emergence of both oxygenases and other oxidoreductases, thus suggesting a wider availability of oxygen around 3.1 Ga.

Oxygenation of the atmosphere had tremendous consequences for the geochemistry and biology of our planet. Molecular oxygen (oxygen hereafter) was introduced by life, primarily by oxygenic photosynthesis, and gradually accumulated to its current level. Its availability promoted innovation at all levels, including the emergence of ~700 new enzymatic reactions[1]. However, at which stage oxygen began to dominate the biosphere remains unclear.

Growing evidence indicates that oxygenic photosynthesis evolved well before geochemical markers unambiguously indicated O_2 accumulation, namely before the Great Oxidation Event (GOE) that occurred around 2.4 Ga; the interpretation of these records is, however, still debated[2–4]. It has accordingly been suggested that it took hundreds of millions of years for oxygen to saturate the geochemical buffers and to start accumulating[5]. What remains unclear therefore is the history of O_2 between the advent of life and the GOE. The available geochemical data led to three alternative hypotheses, all finding both enthusiasts and adversaries. Here we examine these three hypotheses in light of the evolutionary history of enzymes that produce or utilize oxygen (O_2-enzymes), with the notion that O_2 must be present for the emergence of an O_2-utilizing enzyme.

According to the first hypothesis, oxygen was initially produced by a few photosynthetic organisms and became widely available to life only after its accumulation in the atmosphere—that is, across the GOE interval. By this hypothesis, the majority of O_2-enzymes would have emerged during or after the GOE.

The second hypothesis posits that, although biogenic oxygen was insufficient to trigger the geological and geochemical proxies whose signals appear at or near the GOE, oxygen was nonetheless available in the biosphere, either locally or transiently. The former regards oxygen oases, niches occupied by both primary producers (by oxygenic photosynthesis) and consumers[6]. The latter, dubbed oxygen whiffs, relates to periods during which oxygen production by oceanic photosynthesizers was sufficiently high to allow its accumulation in the proximity of their habitat[7,8]. If either of these scenarios is valid, we would expect to see multiple O_2-enzyme emergences before the GOE; foremost, the emergence of O_2-enzymes should coincide with the emergence of oxidative photosynthesis rather than with the GOE.

Finally, in a third scenario, regardless of the environment, oxygen was present in living cells from the dawn of life. Indeed, O_2-enzymes can be identified in strict anaerobes[9–12]. Abiotically produced reactive oxygen species (ROS) such as hydrogen peroxide (H_2O_2) also could have been the primordial source of molecular oxygen[13], and enzymes capable of producing oxygen from such molecules, such as catalase or chlorite lyase, were previously assigned to the last universal common ancestor (LUCA)[14,15]. If oxygen was always available to organisms and at levels sufficient to trigger the emergence of enzymes that utilize it, one would expect to see a notable presence of O_2-enzymes in LUCA and gradual expansion of such enzymes from LUCA thereafter, at a regular pace, as with any other enzyme class.

Given these conflicting hypotheses, we aim to promote understanding of the evolution of O_2-enzymes, and specifically to date their emergence. In doing so, we provide new insight, independent of the geological record, in support of early, pre-GOE, oxygen-dependent metabolism.

Results

Identifying the emergence of O_2-enzyme families. Our analysis workflow is summarized in Fig. 1a. In the first instance, searching through all metabolic reactions in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[16] we identified all known O_2-enzyme families, defined as enzymes that either utilize or produce oxygen. For every such reaction, we extracted the enzyme(s) known to catalyse it, identified its catalytic domain and classified these catalytic domains into 134 protein families (using Pfam classifications; Methods). In most families (81/134), because oxygen-utilizing or -producing activity is assigned to nearly all family members, we could safely assign members of this family as O_2-enzymes (Supplementary Table 1). In these cases we could also assume that the family ancestor was indeed an O_2-utilizing or -producing enzyme (O_2 as the founding function was further validated by examining the Pfam Clans; see below). However, in 53 families the O_2 function is sporadic with no consistent sequence signature distinguishing family members that are O_2-enzymes from those that are not. In fact, in most of these 'niche' families the ancestral function appears unrelated to oxygen. Thus, the emergence of such a family does not necessarily indicate O_2

Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel. ✉e-mail: dan.tawfik@weizmann.ac.il

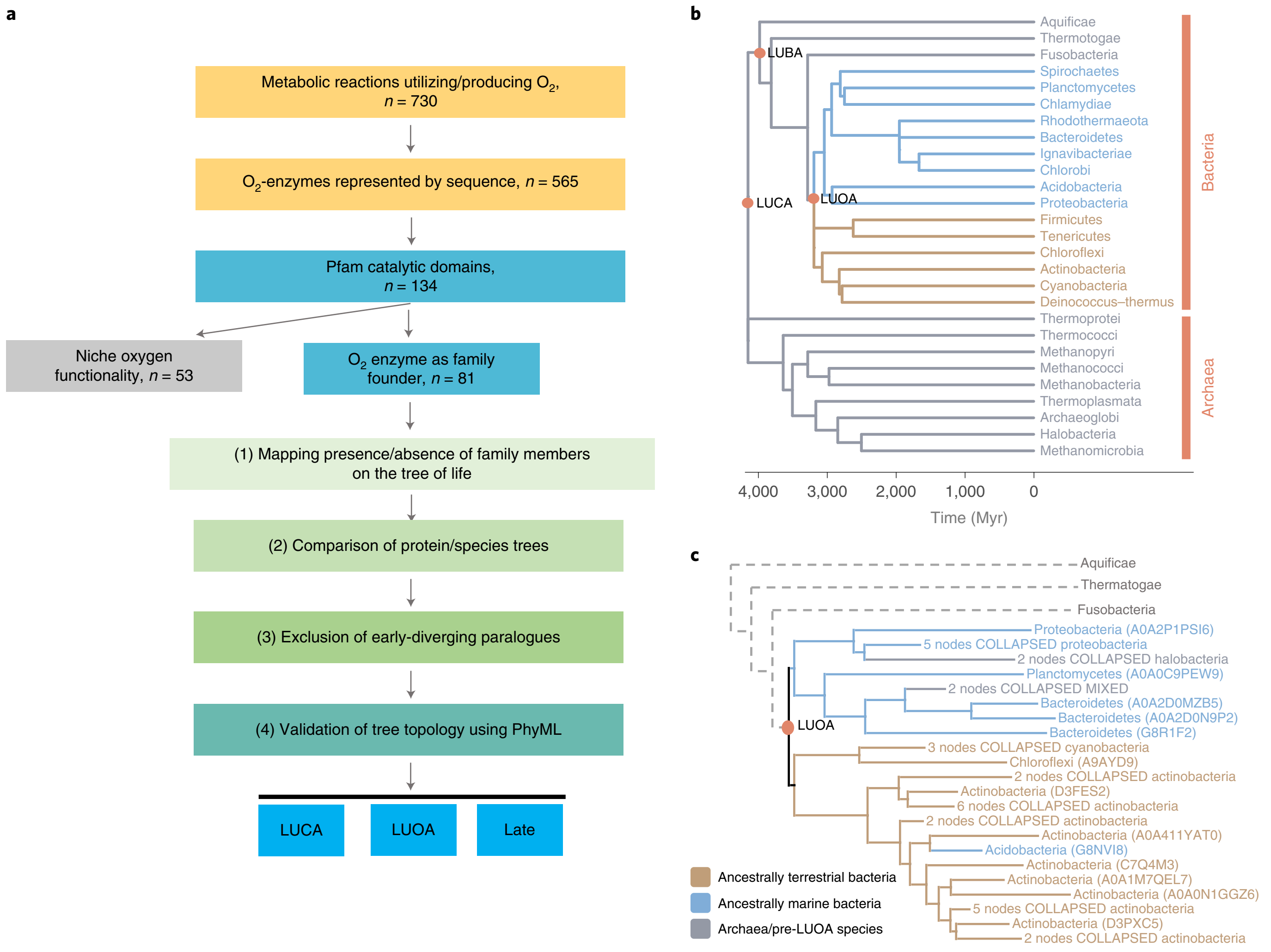
a
Metabolic reactions utilizing/producing O_2, $n = 730$
O_2-enzymes represented by sequence, $n = 565$
Pfam catalytic domains, $n = 134$
Niche oxygen functionality, $n = 53$
O_2 enzyme as family founder, $n = 81$
(1) Mapping presence/absence of family members on the tree of life
(2) Comparison of protein/species trees
(3) Exclusion of early-diverging paralogues
(4) Validation of tree topology using PhyML
LUCA
LUOA
Late
b
LUBA
LUCA
LUOA
Aquificae
Thermotogae
Fusobacteria
Spirochaetes
Planctomycetes
Chlamydiae
Rhodothermaeota
Bacteroidetes
Ignavibacteriae
Chlorobi
Acidobacteria
Proteobacteria
Firmicutes
Tenericutes
Chloroflexi
Actinobacteria
Cyanobacteria
Deinococcus–thermus
Thermoprotei
Thermococci
Methanopyri
Methanococci
Methanobacteria
Thermoplasmata
Archaeoglobi
Halobacteria
Methanomicrobia
Bacteria
Archaea
4,000
3,000
2,000
1,000
0
Time (Myr)
c
Aquificae
Thermatogae
Fusobacteria
Proteobacteria (A0A2P1PSI6)
5 nodes COLLAPSED proteobacteria
2 nodes COLLAPSED halobacteria
Planctomycetes (A0A0C9PEW9)
2 nodes COLLAPSED MIXED
Bacteroidetes (A0A2D0MZB5)
Bacteroidetes (A0A2D0N9P2)
Bacteroidetes (G8R1F2)
LUOA
3 nodes COLLAPSED cyanobacteria
Chloroflexi (A9AYD9)
2 nodes COLLAPSED actinobacteria
Actinobacteria (D3FES2)
6 nodes COLLAPSED actinobacteria
2 nodes COLLAPSED actinobacteria
Actinobacteria (A0A411YAT0)
Acidobacteria (G8NVI8)
Actinobacteria (C7Q4M3)
Actinobacteria (A0A1M7QEL7)
Actinobacteria (A0A0N1GGZ6)
5 nodes COLLAPSED actinobacteria
Actinobacteria (D3PXC5)
2 nodes COLLAPSED actinobacteria
■ Ancestrally terrestrial bacteria
■ Ancestrally marine bacteria
■ Archaea/pre-LUOA species

Fig. 1 | Dating the emergence of O_2-enzymes. a, Dating began with the identification of O_2-enzymes that have known sequences (yellow), followed by mapping of their catalytic domains into families (Pfam families, blue). Next, families that include non-O_2-enzymes were excluded (light grey). The remaining 81 families were subjected to phylogenetic assignment of emergence in four stages: (1) preliminary mapping of the node of emergence based on presence and absence along the prokaryote tree of life (**b**); (2) generation of the protein tree and its comparison to the species tree as demonstrated in **c**; (3) filtering out of paralogues; and (4) validation of the tree topology with an alternative tree-building programme (PhyML). The O_2-enzyme families that passed these four steps were assigned, according to their node of emergence, to either LUCA, LUOA or later than LUOA (Late). Those that failed any one of these steps were left as Unassigned. **b**, Schematic representation of the prokaryote species tree taken from the Time Tree of Life[20]. The last universal common ancestor (LUCA) comprises the root of the tree, with the first split leading to either the last universal bacterial ancestor (LUBA) or to Archaea. The next node, LUOA, corresponds to the separation of terrestrial and marine bacteria. Enzyme families dated to LUOA have, in principle, members in all phyla that diverged from it (blue and brown lines) yet are absent in archaea and the three bacterial phyla that diverged before LUOA (grey lines). **c**, Representative protein tree for a LUOA-assigned enzyme, cysteine dioxygenase (PF05995). The collapsed clades represent those of related sequences, the majority—or all—of which belong to the designated phylum (Methods). Dashed lines represent phyla whose reference genomes have no members of this family. The protein tree largely follows the species tree, but with several inconsistencies arising from horizontal gene transfer—for example, acidobacterial cysteine dioxygenases that cluster with actinobacterial ones, or halobacteria, an archaeal species that acquired its cysteine dioxygenases from bacteria.

availability. Unfortunately, some niche families include enzymes that serve as primary indicators for oxygen utilization, including haem-copper oxygen reductases (HCOs, PF00115). However, the phylogenetic separation of HCOs from nitric oxide reductases that are non-O_2 enzymes is unclear[17], and we also found their protein and species trees to be inconsistent. Other niche families could serve as proxies for the emergence of biogenic oxygen, foremost the L and M proteins of Photosystem II; however, members of this family are often found in phototrophic organisms that do not perform oxygenic photosynthesis[18]. Thus, despite their potential insight, we excluded these families to ensure rigorous assignment (Supplementary Table 2).

Mapping across the prokaryote Tree of Life. Next, we attempted to date the emergence of the 81 families where oxygen utilization or production appears to be the founding, ancestral function (Fig. 1a, steps in green rectangles; the 81 families and their dating details are featured in Supplementary Table 1). To this end, we searched 738 proteomes representing all prokaryotic taxonomic groups for members of these 81 families. We thus mapped the presence of these families across the prokaryote Tree of Life, a distilled version of which is given in Fig. 1b. Seventy-one out of 81 families are consistently present in species that diverged before the GOE (Table 1, step 1). About half are consistently present in both bacteria and archaea and could, in principle, have emerged in LUCA ($n = 35$). Nevertheless, an equally

Table 1 | Emergence of O_2-enzymes along evolutionary time

Assignment	(1) Presence/absence (*n*)	(2) Protein/species tree comparison (*n*)	(3) Paralogues and Pfam Clan assignment (*n*)	(4) PhyML tree topology (*n*)
LUCA	35	3	2	2
LUOA	36	28	23	22
Late	10	12	12	12
Unassigned	0	38	44	45

large number of families ($n = 36$) mapped to the node relating to the major split between terrestrial and marine bacteria[19] and, according to TimeTree[20], date to around 3.1 Ga (±0.15 Gyr; dating is further discussed below). Archaea do not possess these enzymes, nor are they present in the earliest-diverging bacterial clades—the hyperthermophilic Aquificae, Thermatogae and the obligately anaerobic Fusobacteria[19,21] (Fig. 1b). We dubbed this node the Last Universal Oxygen Ancestor (LUOA). The remaining families, denoted as Late ($n = 10$; Table 1), either show late emergence (post-GOE, including eukaryotic and viral origin) or are clade specific, such as ammonia monooxygenase, that is present only in ammonia-oxidizing bacterial taxa. Table 1 shows a summary of the results of the four-stage assignment, indicating the number of O_2-enzyme families whose emergence was assigned to LUCA, LUOA and Late.

Protein trees and the identification of emergence at LUOA. The presence/absence of a protein across the Tree of Life is insufficient to assign its evolutionary history. Proteins are subject to non-vertical evolutionary events such as horizontal gene transfer (HGT), as well as gene losses. The latter would result in more recent assignment compared to the actual emergence and thus, if anything, to underestimation of pre-GOE emergence. HGT, on the other hand, could result in incorrect assignment to an earlier node. To account for HGT, we compared protein family trees to the prokaryote Tree of Life (the species tree; Fig. 1a, step 2). Sequences belonging to the 81 O_2-enzymes families were extracted from 352 reference proteomes (a subset of the proteomes used for mapping of presence/absence) and aligned, and protein trees were inferred using FastTree. The protein trees were rooted and reconciled to maximize matching with the species tree while maintaining statistical confidence[22]. Next, the protein and species trees were compared. HGT is extensive in prokaryotes, and no single protein gives a tree that fully matches the species tree[23,24]. We thus focused on the splits that underlie LUCA (the split between bacteria and archaea) and LUOA (the split between marine and terrestrial bacteria) while ignoring discrepancies in relatively recent lineages (a representative tree is shown in Fig. 1c; all trees are shown in Supplementary Fig. 1). In agreement with the extensive HGT between archaea and bacteria seen in previous analyses[25,26], species–protein tree comparison resulted in validation of only three out of the 35 families initially assigned to LUCA. In contrast, 28 LUOA families were confirmed (Table 1, step 2).

Further validation steps. Two additional steps were taken to support the assignment (steps 3 and 4, Fig. 1a). First, we inspected paralogues to exclude cases where an early duplication biased the tree topology toward a marine–terrestrial split (or bacterial–archaeal in the case of LUCA). To this end, we identified paralogues that may have diverged close to the root (Supplementary Table 3). Twelve families were identified where a relatively large number of phyla contained early-diverging paralogues. For these, trees were rebuilt based on alignments from which these paralogues were removed, as exemplified in Fig. 2. These trees were largely congruent with those containing all sequences: namely, the first bifurcation corresponded to either the marine–terrestrial split (for LUOA, 25 out of 28 families confirmed) or archaeal–bacterial split (3/3 LUCA families confirmed; Table 1 and Supplementary Fig. 2).

Related to the paralogue issue is the fact that Pfam families do not necessarily relate to independent evolutionary emergences. The latter are assigned with higher confidence at the higher hierarchical level, Clans, that typically combine several paralogous families. We thus ensured that not only the family, but also the Pfam Clan to which it belongs, relates to an O_2-enzyme (Supplementary Table 1 and Methods; where this was not the case, we applied additional searches and excluded families to which non-O_2 enzyme sequences could infiltrate). This step of filtering further discarded one LUCA and two LUOA families. The latter two belong to the Cupin and Rossmann Clans that include non-O_2-enzyme families members of which could infiltrate as O_2-enzymes. Similarly, protein trees of LigB that was previously assigned to LUCA[27] indicated archaeal–bacterial as the first split, also after exclusion of paralogues (Supplementary Fig. 2, PF02900). However, the Pfam Clan includes a large family (Memo) whose function is unknown and in-depth study of this Clan suggested that LigB diverged well after LUCA, possibly with the appearance of oxygen[28].

Finally, to validate the protein tree topologies, we examined families that were assigned as either LUCA or LUOA by the previous steps with an alternative method for building phylogenetic trees, PhyML[29] (Fig. 2c and Supplementary Fig. 2).

A burst of emergence of O_2-enzymes and oxidoreductases. The above-described process resulted in 22 families of O_2-enzymes assigned to LUOA and two families to LUCA. Twelve families were now assigned as Late, while the remaining 45 families could not be reliably dated (Table 1, step 4, Unassigned). Thus, according to our dating, the number of O_2-utilizing enzymes has expanded ~12-fold from LUCA to LUOA (from 2 to 24, cumulatively). Furthermore, no emergences of O_2-enzymes were detected at either the last universal archaeal ancestor or the bacterial one (LUBA). It appears, therefore, that a burst of emergence of O_2-enzymes occurred around 3.1 Ga at separation of the ancestral marine and terrestrial bacterial clades.

However, new enzymes have emerged throughout evolutionary time and all enzyme types have expanded during this period (LUCA to LUOA). Did the expansion of O_2-utilizing enzymes exceed that of all other enzyme types, thereby indicating an extraordinary event? To address this question, we identified all Pfam families in the set of proteomes analysed, mapped them to key early nodes and assigned their enzyme classification (EC numbers; Supplementary Table 4). This analysis indicated an expansion twice as large in oxidoreductases between LUBA (the last universal bacterial ancestor) and LUOA compared to all other enzyme classes (Table 2). Notably, LUBA shows roughly the same expansion rate compared to LUCA for all six classes, including oxidoreductases. This unusual expansion of oxidoreductases in general, and specifically of O_2-utilizing enzymes, suggests a dramatic environmental change around the period of LUOA.

Discussion

In summary, our results indicate that the majority of O_2-enzyme families (24/36), the emergence of which could be reliably dated, emerged well before the GOE (Table 1). Even if families that could not be reliably dated (unassigned families) were assigned as post-GOE, about one-third of O_2-enzyme families (24/81) would still relate to the pre-GOE period. The key observation is a burst of emergence of O_2-enzymes around 3.1 Ga. As discussed below, this burst predates the GOE by around 0.5 Gyr and largely coincides

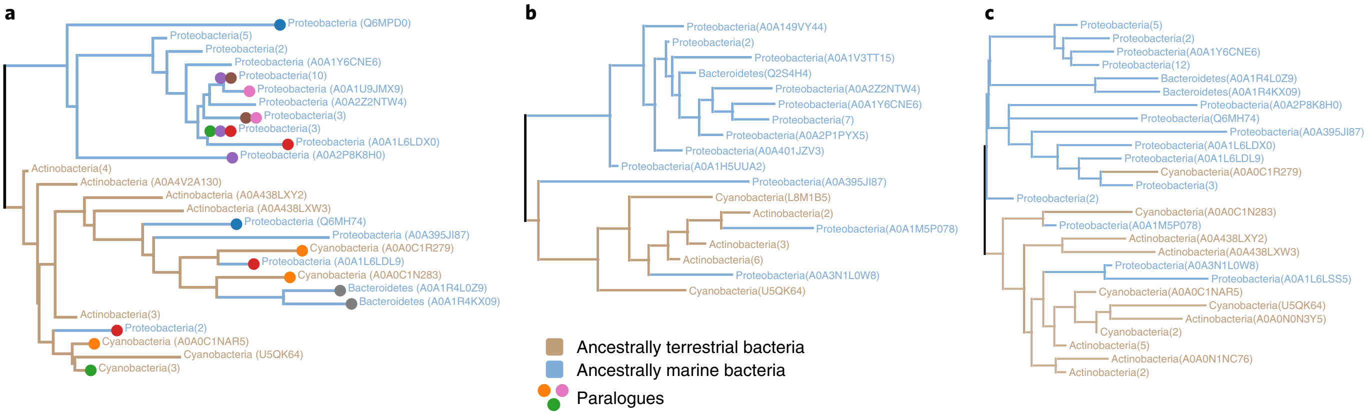
a
Proteobacteria (Q6MPD0)
Proteobacteria(5)
Proteobacteria(2)
Proteobacteria (A0A1Y6CNE6)
Proteobacteria(10)
Proteobacteria (A0A1U9JMX9)
Proteobacteria (A0A2Z2NTW4)
Proteobacteria(3)
Proteobacteria(3)
Proteobacteria (A0A1L6LDX0)
Proteobacteria (A0A2P8K8H0)
Actinobacteria(4)
Actinobacteria (A0A4V2A130)
Actinobacteria (A0A438LXY2)
Actinobacteria (A0A438LXW3)
Proteobacteria (Q6MH74)
Proteobacteria (A0A395JI87)
Cyanobacteria (A0A0C1R279)
Proteobacteria (A0A1L6LDL9)
Cyanobacteria (A0A0C1N283)
Bacteroidetes (A0A1R4L0Z9)
Bacteroidetes (A0A1R4KX09)
Actinobacteria(3)
Proteobacteria(2)
Cyanobacteria (A0A0C1NAR5)
Cyanobacteria (U5QK64)
Cyanobacteria(3)
b
Proteobacteria(A0A149VY44)
Proteobacteria(2)
Proteobacteria(A0A1V3TT15)
Bacteroidetes(Q2S4H4)
Proteobacteria(A0A2Z2NTW4)
Proteobacteria(A0A1Y6CNE6)
Proteobacteria(7)
Proteobacteria(A0A2P1PYX5)
Proteobacteria(A0A401JZV3)
Proteobacteria(A0A1H5UUA2)
Proteobacteria(A0A395JI87)
Cyanobacteria(L8M1B5)
Actinobacteria(2)
Proteobacteria(A0A1M5P078)
Actinobacteria(3)
Actinobacteria(6)
Proteobacteria(A0A3N1L0W8)
Cyanobacteria(U5QK64)
Ancestrally terrestrial bacteria
Ancestrally marine bacteria
Paralogues
c
Proteobacteria(5)
Proteobacteria(2)
Proteobacteria(A0A1Y6CNE6)
Proteobacteria(12)
Bacteroidetes(A0A1R4L0Z9)
Bacteroidetes(A0A1R4KX09)
Proteobacteria(A0A2P8K8H0)
Proteobacteria(Q6MH74)
Proteobacteria(A0A395JI87)
Proteobacteria(A0A1L6LDX0)
Proteobacteria(A0A1L6LDL9)
Cyanobacteria(A0A0C1R279)
Proteobacteria(3)
Proteobacteria(2)
Cyanobacteria(A0A0C1N283)
Proteobacteria(A0A1M5P078)
Actinobacteria(A0A438LXY2)
Actinobacteria(A0A438LXW3)
Proteobacteria(A0A3N1L0W8)
Proteobacteria(A0A1L6LSS5)
Cyanobacteria(A0A0C1NAR5)
Cyanobacteria(U5QK64)
Actinobacteria(A0A0N0N3Y5)
Cyanobacteria(2)
Actinobacteria(5)
Actinobacteria(A0A0N1NC76)
Actinobacteria(2)

Fig. 2 | An example of the LUOA filtering process for family PF08007, Cupin_4. a, In families whose protein trees were consistent with the species tree (Fig. 1c), all paralogues were identified and early-diverging examples were marked on the protein tree generated in step 2 (the tree of all identified family members). **b**, Step 3: paralogues that span across, or are close to, the LUOA node of interest were removed (coloured dots), the remaining sequences were realigned and the tree was rebuilt (Methods). **c**, Step 4: the tree was rebuilt from an alignment generated after removal of paralogues using PhyML rather than FastTree, which was used in steps 2 and 3. Steps 3 and 4 are shown in Fig. 1a. The topology of all three trees is identical with respect to the earliest node corresponding to the split between terrestrial and marine bacteria, thus allowing the emergence of this family to be assigned to LUOA (trees for all other assigned families are provided in Supplementary Fig. 2).

with the emergence of oxygenic photosynthesis[30] (dating error margins are discussed in Supplementary File). However, what we designate as the node of emergence, namely LUOA, probably represents not a single organism, or species, but rather an ensemble of organisms living at that time. Among these were oxygen producers, by oxygenic photosynthesis and possibly by other means, as well as users whose presence is indicated by the emergency of multiple O_2-utilizing enzymes and numerous other oxidoreductases. Indeed, the early Tree of Life for prokaryotes might be conceived as a web of life rather than a set of discretely defined species[23].

The LUOA O_2-enzymes. A wide range of enzymes that utilize oxygen appear to have emerged in LUOA, foremost being those involved in the degradation of fatty acids, sterols and, especially, of aromatic amino acids. The reactivity of oxygen is the key to breaking down the highly stable carbon–carbon bonds of these compounds. It has also been suggested that the appearance of oxygen triggered the systematic incorporation of aromatic amino acids into proteins[31], possibly as a means of protecting them from ROS[32]. The LUOA origin of enzymes that degrade aromatic amino acids, and also of degradation pathways, as discussed below, supports the hypothesis that aromatic amino acids, and tryptophan in particular, were the latest addition to the genetic code[33], possibly post LUCA[34]. Also notable in this context are two lysine-degrading enzyme families, because lysine is also considered to have emerged late. While the function of the LUOA enzymes seems to have some common denominator, their active site chemistries are diverse, including both haem and non-haem enzymes, with the latter including iron, copper and nickel enzymes (Supplementary Table 1).

Oxygen availability is reflected not only through individual O_2-enzymes but also by the emergence of oxygen-dependent metabolic pathways[1]. Indeed, following examination of all metabolic pathways containing two or more O_2-enzymes belonging to any of the 81 families analysed, we identified 21 pathways with at least one LUOA enzyme (Supplementary Table 5). As expected, we identified tryptophan, phenylalanine and tyrosine degradation pathways with at least one oxygen-dependent step being catalysed by a LUOA enzyme, thus advocating for LUOA emergence of these pathways. Furthermore, while the LUOA O_2-enzymes are detected individually in our set of prokaryotic genomes (that is, one family representative per genome), in organisms that diverged later (metazoan or even Mammalia) they appear as multiple paralogues that catalyse as many as three steps in the same pathway. Thus, over time, the ancestral LUOA enzymes gave rise to multiple families via gene duplications.

Finally, O_2-enzymes emerging before the GOE probably had to operate at low oxygen levels. While the K_M values (the Michaelis constant, a proxy of enzyme affinity for O_2) at the time of emergence are unknown, multiple evidences indicate that O_2-enzymes can operate at very low oxygen concentrations. For example, O_2-enzymes that mediate steroid biosynthesis operate at nanomolar dissolved oxygen concentrations indicating that these enzymes could have emerged before the GOE[35]. Furthermore, O_2-enzymes are far more abundant in anaerobes than expected, including in strict anaerobes; some of these enzymes are involved in detoxification of ROS, but most are O_2-utilizing enzymes, including enzymes likely to be essential such as pyridoxal 5'-phosphate synthase[9]. Finally, the BRENDA database, which lists all known kinetic parameters, indicates a median K_M value for oxygen of ~54 μM, about one-fifth of the saturating aqueous oxygen concentration below the present atmospheric level, and values of 2.5 μM or lower (≤1% of the saturating concentration) have also been recorded (Supplementary Fig. 3).

The origins of biogenic oxygen. LUCA is widely deemed a strict anaerobe. A recent assignment of the LUCA protein families under this assumption indicated, as expected, only a few O_2-enzyme families[36]. Our analysis, which was not biased against aerobes as such, also yielded only two LUCA O_2-enzyme families. Both the previous assignments and ours are in agreement with the abundance of anaerobes in archaea and in early-diverging bacterial species. However, the assigned families differ because, in general, current LUCA assignments vary widely. Indeed, many LUCA assignments have been questioned[37] and, accordingly, families previously assigned to LUCA were annotated by us as "Unassigned" due to inconsistencies between species and protein trees and/or to infiltration of non-O_2-enzyme sequences (for example, LigB that was previously annotated as LUCA[36]). Nonetheless, the two families assigned here to LUCA were previously noted as such[14,27,38]. Iron-manganese superoxide dismutase is an ROS detoxifying enzyme that produces oxygen. This enzyme is highly abundant in anaerobes, as are other

Table 2 | The preferential expansion of oxidoreductases at LUOA

EC class(es)	EC name	LUCA	LUBA	LUOA	LUBA/LUCA	LUOA/LUBA
1–6	All classes	172	219	437	1.27	1.99
2	Transferases	66	85	138	1.29	1.62
3	Hydrolases	36	51	121	1.42	2.37
4	Lyases	18	19	35	1.06	1.84
5	Isomerases	16	19	34	1.19	1.8
6	Ligases	32	37	46	1.16	1.24
Average of classes 2–6	–	–	–	–	1.22 (±0.14)	1.77 (±0.4)
1	Oxidoreductases	24	32	109	1.33	3.41

The number of Pfam enzymes families assigned to LUCA, LUBA and LUOA, categorized by their enzyme classes. The data provided are cumulative: namely, the families in LUBA include those in LUCA and those in LUOA include those in both LUBA and LUCA.

ROS-neutralizing enzymes that produce O_2, such as catalases and peroxidases[9] (many of which are assigned to LUCA by the preliminary criteria of presence/absence; Supplementary Table 1). The second LUCA-assigned enzyme is cytochrome bd oxidase. Although usually associated with the respiratory chain, it is widely present in anaerobes, probably as protection against oxidative stress[11,39,40]. The identification of these two enzymes therefore supports the notion of intracellular oxygen production in LUCA.

Early oxygen availability is also indicated by the phylogenetic dating of oxygenic photosynthesis, the key source of molecular oxygen on Earth. Molecular clock studies yielded age estimates for the origin of cyanobacteria, the earliest photosynthesizers, that span widely from 3.5 to ~2 Ga (refs. [41–43]). Specifically, a relaxed molecular clock of the Type II reaction centre proteins dates the last common ancestor of the oxygenic system at 3.1 Ga (ref. [30]) (discussed further in the Supplementary File). We undertook an alternative approach to dating enzymes that utilize/produce oxygen and therefore indicate oxygen availability. Furthermore, we did not focus on cyanobacteria but instead examined the prokaryote species tree in its entirety. Nonetheless, our study also indicates that oxygen became available to biology around 3.1 Ga.

We note, however, that like the geochemical evidence discussed below, molecular data—including ours—are heavily dependent on the sequence datasets used, the assumptions made regarding the Tree of Life and the specific model calibrations[44].

Can geochemical and phylogenetic indications be reconciled? Geochemical estimates regarding the timing of biological oxygen production vary, although it is generally accepted that oxygenation of the atmosphere occurred between 2.5 and 2.3 Ga (refs. [45–47]). Our phylogenetic analysis consistently shows a burst of emergence of O_2-enzymes at around 3.1 Ga (±0.1 Gyr), long before the atmospheric rise in oxygen, as calibrated by the plausible separation of the major marine and terrestrial bacterial phyla[19]. However, our dating is not necessarily contradictory to geochemical observations from rock records. For instance, a recent study interpreted uranium isotope signatures as evidence for biological oxygen production at around 3 Ga (ref. [48]). Similar inferences have been drawn using molybdenum, chromium and iron isotopes present in palaeosols formed ≥2.95 Ga (refs. [49–51]), although the interpretation of such records continues to be debated[52]. Nonetheless, at present, given multiple independent indications, both geochemical[53] and phylogenetic, it seems highly likely that oxygen was available to life at a relatively early stage, at least 0.5 Gyr before the GOE.

Methods

Identification of O_2-enzyme families. Our analysis workflow is summarized in Fig. 1a. To identify all known O_2-enzyme families, we searched the KEGG database[16] for all reactions involving dioxygen as either substrate (in nearly all cases) or product (for example, catalase). We also identified enzymes with H_2O_2 as a substrate (peroxidases), since this is a product of an oxygen-based metabolism. For every such reaction, we extracted the enzyme(s) known to catalyse them. Overall, we identified 730 enzymes, mostly belonging to two classes: dioxygenases (EC 1.13.-.-) and monooxygenases (1.14.-.-). We assigned the Pfam family corresponding to the enzyme's catalytic domain (as opposed to other auxiliary domains that do not relate to O_2) by comparing structures and identifying the domain present in all related enzymes. For EC classes where multiple non-redundant structures were not available, we ran HMMsearch[54] on ExPASy Enzyme sequences[55] against the Pfam HMM database with a Pfam-gathering threshold, and identified the catalytic domain by virtue of it being shared by all related enzymes. Finally, we cross-validated the EC-to-Pfam mappings using ECDomainMiner[56]. Overall, we mapped 565/730 of the O_2-enzymes to 134 Pfam families (a given EC number—that is, enzyme class—may be catalysed by multiple families and, conversely, a given Pfam family may include dozens of EC numbers; Supplementary Tables 1 and 2). The remaining 165 enzymes were either not represented by a sequence in the ExPASy Enzyme database or a reliable assignment of their catalytic domain was not feasible.

Identification of niche versus founding functionalities. In most families, oxygen-utilizing or -producing activity is assigned to nearly all family members. We could thus safely assign members of this family as O_2-enzymes, and also assume that the family ancestor was indeed a utilizer or producer of O_2 (O_2 founding function). However, in some families, non-O_2 enzyme functions are dominant; in this case, family members that are O_2-enzymes cannot be distinguished from those that are not. Primarily, the emergence of such a family does not indicate O_2 availability. We therefore identified the non-O_2-associated EC classes included in the mapped Pfam families using the ECDomainMiner database[56]. Following manual inspection of the EC classes in each family, we classified 81 Pfam families with O_2 founding function and 53 where oxygen activity is a niche function (Supplementary Tables 1 and 2, respectively). A typical example is shown by families comprising enzymes that perform the same or similar reactions with alternative electron acceptors—principally dehydrogenases that use $NAD(P)^+$ that are in the same Pfam family as oxidases that use O_2 (ref. [57])—for example, vanillyl alcohol oxidase and 4-methylphenol dehydrogenase that share high sequence identity (>30%)[58]. Since the vast majority of family members are NAD(P) dependent, the ancestor was probably a dehydrogenase rather than an O_2-dependent enzyme. Accordingly, O_2 niche families were excluded from our analysis.

Presence/absence across the prokaryote Tree of Life. In the first instance, the presence/absence of members of the founding O_2-enzyme families was mapped onto the prokaryote Tree of Life. Of the currently available 6,800 reference prokaryotic proteomes in UniProt, 738 representative species spanning all taxonomic prokaryote families in the TimeTree database were selected[20] (Supplementary Table 6). Using the hidden Markov models (HMM) sequence profiles provided by Pfam, homologues of all founding families were identified using hmmsearch with the Pfam-defined gathering e-value threshold across all selected species. A list of phyla encoding enzymes from all founding families can be found in Supplementary Table 1.

Phylogenetic trees. The comparison of protein and species trees demands manual curation. We therefore selected a subset of 352 reference proteomes representing all taxonomic orders (a subset of the proteomes used for mapping presence/absence across the prokaryote Tree of Life). For each founding O_2-enzyme Pfam family, the sequences obtained from HMM searches on this subset of proteomes were aligned using Mafft software with the *-linsi* option[59]. The alignments were trimmed using trimaAl[60] (with *-gappyout* option) to remove positions where gaps dominate. For each alignment, the phylogenetic tree was constructed using FastTree[61] (with *-pseudo -spr* 4 *-mlacc* 2 *-slownni* parameters and the Jones–Taylor–

Thornton (JTT) evolutionary model). The trees were reconciled, and also with the TimeTree-derived species tree topology, using the treefix-DTL algorithm[22] that minimizes the effects of gene duplications, transfers and losses. Next, the trees were rooted using the most optimal root as selected by the MAD programme[62] and compared to the species tree (Fig. 1a, step 2). To this end, clades were consolidated and their dominating phylum was assigned. Consolidation was performed by systematically examining the average distance to leaves, consolidating all clades where this distance was shorter than the average length across all tree edges. The phylum whose sequences constituted the majority (>50%) of the given clade was assigned as the representative of the collapsed clade; in the absence of a clear majority, collapsed clades were assigned as 'mixed' (a representative collapsed tree is shown in Fig. 1b, and all trees are provided in Supplementary Fig. 1).

For putative LUCA and LUOA families, we also inspected the paralogous sequences (step 3). For every such family, we identified all homologous sequences encoded in the same genome and annotated them as paralogous. We then determined the average number of paralogues per phylum. To identify early-diverging paralogues, the average pair-wise sequence identity between paralogues was compared to the average identity between sequences belonging to marine and terrestrial clades (for LUOA families) and between the archaeal and bacterial clades (for LUCA families; Supplementary Table 3). If paralogue identity was in the same range or lower than that between terrestrial and marine sequences (LUOA), or bacterial and archaeal sequences (LUCA), the paralogues were annotated as early diverging (that is, if paralogues were as diverged as the earliest split). Additionally, the fractions of phyla and species affected by the presence of paralogues were calculated. Subsequently, for LUOA families where ≥30% of phyla contained an extensive number of paralogues we removed all early-diverging paralogues, revised the phylogenetic trees and compared them to the tree of all sequences (Supplementary Fig. 2).

In yet another filtering step, Pfam Clans were assigned to each Pfam family. All LUCA and LUOA families were then analysed to examine whether they belong to a Pfam Clan dominated by oxygen-dependent enzymes. For those families where the Pfam Clan is predominantly represented by non-O_2 families, the HMM profiles of all Clan members were retrieved and used as queries for hmmsearch on the database of 352 reference proteomes with the Pfam family-specific gathering e-value threshold. Next, the e-values of sequences assigned to both O_2 and non-O_2 Pfam families were compared, and the lower value was used to assign sequence identity as either O_2-enzyme or non-O_2-enzyme. LUCA and LUOA families with ≥10% of overlapping sequence assignments were classified as Unassigned, because non-O_2-enzyme sequences could have been misassigned as O_2-enzymes.

Finally, in step 4 of the analysis, we built alternative trees for all putative LUCA and LUOA families using PhyML (with default parameters and the JTT evolutionary model). These trees were reconciled and rooted as described above.

LUCA assignment. A family was considered LUCA if its representatives could be identified in ≥50% of bacterial (nine phyla) and ≥50% of archaeal phyla (four phyla), and the protein tree indicated that the oldest split corresponded to the divergence of bacteria and archaea. We further considered a minimum number of bacteria–archaea splits, and relatively large intra-/interdomain distances (D_{avg}), as criteria for LUCA origin[37] (Supplementary Table 7). Although none of these individual criteria was sufficient to confirm or rule out LUCA origin, in combination they allowed three families to be assigned as such with relative confidence (with one family, LigB, ultimately discarded due to the Clan containing a non-O_2-enzyme family).

Analysis of metabolic pathways. We searched the metabolic pathways listed in the MetaCyc database to identify those containing O_2-enzymes that belong to any of the 81 dated families. We identified 312 such pathways and assigned the nodes of emergence for their O_2-enzymes. Of these, 21 contain at least one LUOA enzyme (Supplementary Table 5).

Enzymatic expansion analysis. Using the subset of 352 reference proteomes, we identified all Pfam families present in each proteome using HMMsearch and the set of HMM profiles from the Pfam database (v.32.0) with the Pfam-defined gathering threshold. Next, we assigned Pfam families to LUCA, LUBA or LUOA using the following criteria. LUCA: Pfam families present in ≥90% of bacterial and ≥90% of archaeal phyla; LUBA: families not present in archaea but present across ≥90% of bacterial phyla; LUOA: families present in both the earliest-diverging terrestrial and marine bacterial clades but present neither in archaea nor the outgroup bacterial clades (Aquificae, Thermotogae, Fusobacteria). Next, for every Pfam, we identified all associated EC classes using ECDomainMiner[56]. For Pfam families including multiple enzymatic activities, we chose the most common (based on the first EC digit). If two or more EC classes were equally common within a Pfam family, we retained all of them. The list of LUCA, LUBA and LUOA Pfam families with assigned EC classes can be found in Supplementary Table 4.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequence alignments and phylogenetic trees analysed in this study can be found at https://figshare.com/projects/The_evolution_of_oxygen-utilizing_enzymes_suggests_early_biosphere_oxygenation/93818. All other data generated or analysed during this study are included in the published article (and its Supplementary File).

Received: 3 June 2020; Accepted: 16 December 2020;
Published online: 25 February 2021

References

1. Raymond, J. & Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767 (2006).
2. Holland, H. D. The oxygenation of the atmosphere and oceans. *Philos. Trans. R. Soc. Lond. B* https://doi.org/10.1098/rstb.2006.1838 (2006).
3. Holland, H. D. Volcanic gases, black smokers, and the great oxidation event. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).
4. Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289**, 756–758 (2000).
5. Kasting, J. F. & Howard, M. T. Atmospheric composition and climate on the early Earth. *Philos. Trans. R. Lond. Soc. B* **361**, 1733–1741 (2006).
6. Olson, S. L., Kump, L. R. & Kasting, J. F. Quantifying the areal extent and dissolved oxygen concentrations of Archean oxygen oases. *Chem. Geol.* **362**, 35–43 (2013).
7. Lalonde, S. V. & Konhauser, K. O. Benthic perspective on Earth's oldest evidence for oxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **112**, 995–1000 (2015).
8. Duan, Y. et al. A whiff of oxygen before the great oxidation event? *Science* **317**, 1903–1906 (2007).
9. Jabłońska, J. & Tawfik, D. S. The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype. *Free Radic. Biol. Med.* **140**, 84–92 (2019).
10. Sousa, F. L., Nelson-Sathi, S. & Martin, W. F. One step beyond a ribosome: the ancient anaerobic core. *Biochim. Biophys. Acta Bioenerg.* **1857**, 1027–1038 (2016).
11. Das, A., Silaghi-Dumitrescu, R., Ljungdahl, L. G. & Kurtz, D. M. Cytochrome bd oxidase, oxidative stress, and dioxygen tolerance of the strictly anaerobic bacterium *Moorella thermoacetica*. *J. Bacteriol.* **187**, 2020–2029 (2005).
12. Ettwig, K. F. et al. Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* **464**, 543–548 (2010).
13. Slesak, I., Slesak, H. & Kruk, J. Oxygen and hydrogen peroxide in the early evolution of life on earth: in silico comparative analysis of biochemical pathways. *Astrobiology* **12**, 775–784 (2012).
14. Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor – exobiology from a terrestrial perspective. *Res. Microbiol.* **157**, 57–68 (2006).
15. Hofbauer, S., Schaffner, I., Furtmüller, P. G. & Obinger, C. Chlorite dismutases – a heme enzyme family for use in bioremediation and generation of molecular oxygen. *Biotechnol. J.* **9**, 461–473 (2014).
16. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
17. Gribaldo, S., Talla, E. & Brochier-Armanet, C. Evolution of the haem copper oxidases superfamily: a rooting tale. *Trends Biochem. Sci.* **34**, 375–381 (2009).
18. Fischer, W. W., Hemp, J. & Johnson, J. E. Evolution of oxygenic photosynthesis. *Annu. Rev. Earth Planet. Sci.* **44**, 647–683 (2016).
19. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).
20. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
21. Giovannelli, D. et al. Insight into the evolution of microbial metabolism from the deep-branching bacterium, *Thermovibrio ammonificans*. *eLife* **6**, e18990 (2017).
22. Bansal, M. S., Wu, Y.-C., Alm, E. J. & Kellis, M. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* **31**, 1211–1218 (2015).
23. Gribaldo, S. & Brochier, C. Phylogeny of prokaryotes: does it exist and why should we care? *Res. Microbiol.* **160**, 513–521 (2009).
24. Nelson-Sathi, S. et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
25. Fuchsman, C. A., Collins, R. E., Rocap, G. & Brazelton, W. J. Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ.* **2017**, e3865 (2017).
26. Garushyants, S. K., Kazanov, M. D. & Gelfand, M. S. Horizontal gene transfer and genome evolution in *Methanosarcina*. *BMC Evol. Biol.* **15**, 102 (2015).
27. Weiss, M. C. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
28. Maxwell Burroughs, A., Glasner, M. E., Barry, K. P., Taylor, E. A. & Aravind, L. Oxidative opening of the aromatic ring: tracing the natural history of a large

superfamily of dioxygenase domains and their relatives. *J. Biol. Chem.* **294**, 10211–10235 (2019).
29. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
30. Cardona, T., Sánchez-Baracaldo, P., Rutherford, A. W. & Larkum, A. W. Early Archean origin of Photosystem II. *Geobiology* **17**, 127–150 (2019).
31. Granold, M., Hajieva, P., Toşa, M. I., Irimie, F. D. & Moosmann, B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl Acad. Sci. USA* **115**, 41–46 (2018).
32. Gray, H. B. & Winkler, J. R. Living with oxygen. *Acc. Chem. Res.* **51**, 1850–1857 (2018).
33. Fournier, G. P. & Alm, E. J. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *J. Mol. Evol.* **80**, 171–185 (2015).
34. De Pouplana, L. R., Frugier, M., Quinn, C. L. & Schimmel, P. Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. *Proc. Natl Acad. Sci. USA* **93**, 166–170 (1996).
35. Waldbauer, J. R., Newman, D. K. & Summons, R. E. Microaerobic steroid biosynthesis and the molecular fossil record of Archean life. *Proc. Natl Acad. Sci. USA* **108**, 13409–13414 (2011).
36. Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V. & Martin, W. F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet.* **14**, e1007518 (2018).
37. Berkemer, S. J. & McGlynn, S. E. A new analysis of archaea–bacteria domain separation: variable phylogenetic distance and the tempo of early evolution. *Mol. Biol. Evol.* **37**, 2332–2340 (2020).
38. Ślesak, I., Ślesak, H., Zimak-Piekarczyk, P. & Rozpądek, P. Enzymatic antioxidant systems in early anaerobes: theoretical considerations. *Astrobiology* **16**, 348–358 (2016).
39. Juty, N. S., Moshiri, F., Merrick, M., Anthony, C. & Hill, S. The *Klebsiella pneumoniae* cytochrome bd' terminal oxidase complex and its role in microaerobic nitrogen fixation. *Microbiology* **143**, 2673–2683 (1997).
40. Jay, Z. J. et al. Predominant *Acidilobus*-like populations from geothermal environments in Yellowstone National Park exhibit similar metabolic potential in different hypoxic microbial communities. *Appl. Environ. Microbiol.* **80**, 294–305 (2014).
41. Falcón, L. I., Magallón, S. & Castillo, A. Dating the cyanobacterial ancestor of the chloroplast. *ISME J.* **4**, 777–783 (2010).
42. Shih, P. M., Hemp, J., Ward, L. M., Matzke, N. J. & Fischer, W. W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19–29 (2017).
43. Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
44. Magnabosco, C., Moore, K. R., Wolfe, J. M. & Fournier, G. P. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology* **16**, 179–189 (2018).
45. Gumsley, A. P. et al. Timing and tempo of the great oxidation event. *Proc. Natl Acad. Sci. USA* **114**, 1811–1816 (2017).
46. Luo, G. et al. Rapid oxygenation of Earth's atmosphere 2.33 billion years ago. *Sci. Adv.* **2**, e1600134 (2016).
47. Farquhar, J. & Wing, B. A. Multiple sulfur isotopes and the evolution of the atmosphere. *Earth Planet. Sci. Lett.* **213**, 1–13 (2003).
48. Wang, X. et al. A Mesoarchean shift in uranium isotope systematics. *Geochim. Cosmochim. Acta* **238**, 438–452 (2018).
49. Planavsky, N. J. et al. Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat. Geosci.* **7**, 283–286 (2014).
50. Crowe, S. A. et al. Atmospheric oxygenation three billion years ago. *Nature* **501**, 535–538 (2013).
51. Eickmann, B. et al. Isotopic evidence for oxygenated Mesoarchaean shallow oceans. *Nat. Geosci.* **11**, 133–138 (2018).
52. Albut, G. et al. Modern rather than Mesoarchaean oxidative weathering responsible for the heavy stable Cr isotopic signatures of the 2.95 Ga old Ijzermijn iron formation (South Africa). *Geochim. Cosmochim. Acta* **228**, 157–189 (2018).
53. Catling, D. C. & Zahnle, K. J. The Archean atmosphere. *Sci. Adv.* **6**, eaax1420 (2020).
54. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
55. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
56. Alborzi, S. Z., Devignes, M. D. & Ritchie, D. W. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinform.* **18**, 107 (2017).
57. Raymond, J. & Blankenship, R. E. Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology* **2**, 199–203 (2004).
58. Gygli, G., Lucas, M. F., Guallar, V. & van Berkel, W. J. H. The ins and outs of vanillyl alcohol oxidase: identification of ligand migration paths. *PLoS Comput. Biol.* **13**, e1005787 (2017).
59. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
60. 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).
61. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
62. Tria, F. D. K., Landan, G. & Dagan, T. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* **1**, 193 (2017).

Acknowledgements

We thank the Israel Science Foundation for funding (grant nos. 980/14 and 2575/20). This research was partially supported by the Israeli Council for Higher Education via the Weizmann Data Science Research Center and by a research grant from Madame Olga Klein – Astrachan. D.S.T. is the Leon and Nella Benoziyo Professor of Biochemistry. We thank S. Malik, D. Matelska and B. Ross for valuable comments on this manuscript and, especially, I. Halevy and P. Crockford, who inspired us to pursue this study and whose knowledge of palaeogeochemistry guided us throughout.

Author contributions

J.J. and D.S.T. conceived and designed the study, performed the analysis and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-020-01386-9.

Correspondence and requests for materials should be addressed to D.S.T.

Peer review information *Nature Ecology & Evolution* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

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

nature research

Corresponding author(s): Dan S. Tawfik (dan.tawfik@weizmann.ac.il)

Last updated by author(s): Dec 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected from publicly available databases: representative proteomes were downloaded from UniProt, known enzyme sequences from ExPASy, protein families HMM profiles from Pfam, species tree topology from TimeTree, and metabolic networks from MetaCyc. All sequence accession numbers used in the study can be found in the supplementary material.
Data analysis	HMMsearch from HMMER v3.3, Mafft v6.240, trimAl v1.2, TreeFix-DTL v1.0.2, FastTree v2.1, MAD v2.2, An in-house algorithm Oxyphen (https://github.com/wildberry93/oxyphen), The D values and splits calculation with script provided by Sarah Berkemer (https://github.com/bsarah/treeSplits)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences   □ Behavioural & social sciences   ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study aimed at dating the emergence of oxygen enzymes as a proxy for the Earth oxygenation timelines. We consolidated information about known enzymes, protein families and organisms' phylogeny. We constructed phylogenetic trees of oxygen enzymes families using a set of 352 reference proteomes and analyzed the emergence of oxygen-dependent metabolic pathways.
Research sample	Due to the evolutionary time frames the study focuses on, we selected a set of prokaryotic organisms representing all taxonomic orders and families. Eukaryote and viral proteomes were only analyzed for those enzymes absent in prokaryotes.
Sampling strategy	In the study, we used two prokaryotic species sets. The larger dataset that includes representatives of all prokaryotic families (738 reference proteomes) was used to map the presence/absence of enzymes on the tree of life. The smaller set spans representatives of all prokaryotic orders (352 reference proteomes) and was used to compare species/protein trees. The topology of the Tree of Life was retrieved from the Time Tree resource, along with estimated times of divergence of taxa.
Data collection	All data used in the study was taken from publicly available resources such as Uniprot, TimeTree, Pfam, ExPASy, MetaCyc.
Timing and spatial scale	n/a
Data exclusions	We searched for homologs only within a defined set of 738 reference proteomes.
Reproducibility	All the methods, parameters and data used in the study were listed or included in the supplementary material.
Randomization	Randomization is built-in the algorithms used for reconstructing phylogenetic trees.
Blinding	Blinding was not relevant to the study - it's a bioinformatics study.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging