nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0644-x

Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin

Holly C. Betts[1], Mark N. Puttick[1,2], James W. Clark[1], Tom A. Williams[1,3], Philip C. J. Donoghue[1] and Davide Pisani[1,3]*

Establishing a unified timescale for the early evolution of Earth and life is challenging and mired in controversy because of the paucity of fossil evidence, the difficulty of interpreting it and dispute over the deepest branching relationships in the tree of life. Surprisingly, it remains perhaps the only episode in the history of life where literal interpretations of the fossil record hold sway, revised with every new discovery and reinterpretation. We derive a timescale of life, combining a reappraisal of the fossil material with new molecular clock analyses. We find the last universal common ancestor of cellular life to have predated the end of late heavy bombardment (>3.9 billion years ago (Ga)). The crown clades of the two primary divisions of life, Eubacteria and Archaebacteria, emerged much later (<3.4 Ga), relegating the oldest fossil evidence for life to their stem lineages. The Great Oxidation Event significantly predates the origin of modern Cyanobacteria, indicating that oxygenic photosynthesis evolved within the cyanobacterial stem lineage. Modern eukaryotes do not constitute a primary lineage of life and emerged late in Earth's history (<1.84 Ga), falsifying the hypothesis that the Great Oxidation Event facilitated their radiation. The symbiotic origin of mitochondria at 2.053–1.21 Ga reflects a late origin of the total-group Alphaproteobacteria to which the free living ancestor of mitochondria belonged.

Attempts to investigate the emergence of life and its subsequent evolution have traditionally focused on the fossil record. However, this record, especially when looking at the earliest scions of life, is minimal and interpretation is made harder due to difficulties substantiating relationships within the earliest branching lineages of the tree of life[1,2]. Despite its problematic nature, the fossil record remains the main source of information for the timeline of life's evolution. We attempt to shed light on this early period by presenting a molecular timescale based on the ever-growing collection of genetic data, and explicitly incorporating uncertainty associated with fossil sampling, ages and interpretations[1,3–5].

Calibrations are a crucial component of divergence time estimation. Relative divergence times can be inferred using alternative lines of evidence; for example, horizontal gene transfers[6]. However, an absolute timescale for evolutionary history can only be derived when calibrations are included in the analyses[7,8]. We derived a suite of calibrations, following best practice[4] for the fundamental clades within the tree of life, drawing on multiple lines of evidence, including physical fossils, biomarkers and isotope geochemistry[2]. Two key calibrations, for the last universal common ancestor (LUCA) and the oldest total-group eukaryotes, constrain the whole tree by setting a maximum on the root, while also informing the timing of divergence of eukaryotes within Archaea[9,10]. Putative records for life extend back to the Eoarchaean, including microfossils[11,12], stromatolites[13] and isotope data[14,15] from the ~3.8 billion years ago (Ga) Isua Greenstone Belt (Greenland). However, these records have been contested[16–18]. Microfossils from the ~3.4 Ga Strelley Pool Formation, Australia, are the oldest conclusive evidence to constrain the age of LUCA[19]. The fossils, many of which are arranged in chains of cells, have been shown, through nanoscale imaging and Raman spectroscopy, to exhibit a complex morphology with a central, usually hollow, lenticular body and a wall that is either smooth or in some cases reticulated; these features are beyond the scope of pseudofossils[2]. The Strelley Pool Formation also contains other microfossils[20–22], in association with both distinct $\partial^{13}C_{org}$ and $\partial^{13}C_{inorg}$[23] and pyrite indicative of sulfur metabolisms[24], along with stromatolites that exhibit biological structure[25]. Overall, these data allow us to confidently use the Strelley Pool Biota as the oldest, undisputable, record of life. For a maximum constraint on the age of LUCA, we considered the youngest event on Earth that life could not have survived. Conventionally, this is taken as the end of the episode of late heavy bombardment, but modelling has shown that this would not have been violent enough for planet sterilization[26]. However, the last formative stage of Earth's formation—the Moon-forming impact—melted and sterilized the planet. The oldest fossil remains that can be ascribed to crown Eukaryota are ~1.1 Ga *Bangiomorpha pubescens*[27,28], which can be confidently assigned to the red algal total group (Rhodophyta). Older fossil remains from the >1.561 Ga Chitrakoot Formation have been tentatively interpreted as red algae[29]; however, current knowledge of their morphology does not allow for an unequivocal assignment to crown Archaeplastida. The oldest fossil remains that can be ascribed with certainty to total-group Eukaryota are acritarchs from the >1.6191 Ga Changcheng Formation, North China[30], which are discriminated from prokaryotes by their large size (40–250 µm) and complex wall structure, including striations, longitudinal ruptures and a trilaminar organization. However, these structures do not indicate membership of any specific crown eukaryote clade, only allowing us to use these records to minimally constrain the timing of divergence between the Eukaryota and their archaebacterial sister lineage, Asgardarchaeota[9,10,31]. As there is no other evidence to maximally constrain the time of divergence between Eukaryota and Asgardarchaeota, we used the same maximum placed on LUCA; that is, the Moon-forming impact. These key time constraints were combined with nine others (see Supplementary Information) to calibrate a timescale of life estimated from a dataset of 29 highly

[1]School of Earth Sciences, University of Bristol, Bristol, UK. [2]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK. [3]School of Biological Sciences, University of Bristol, Bristol, UK. *e-mail: davide.pisani@bristol.ac.uk

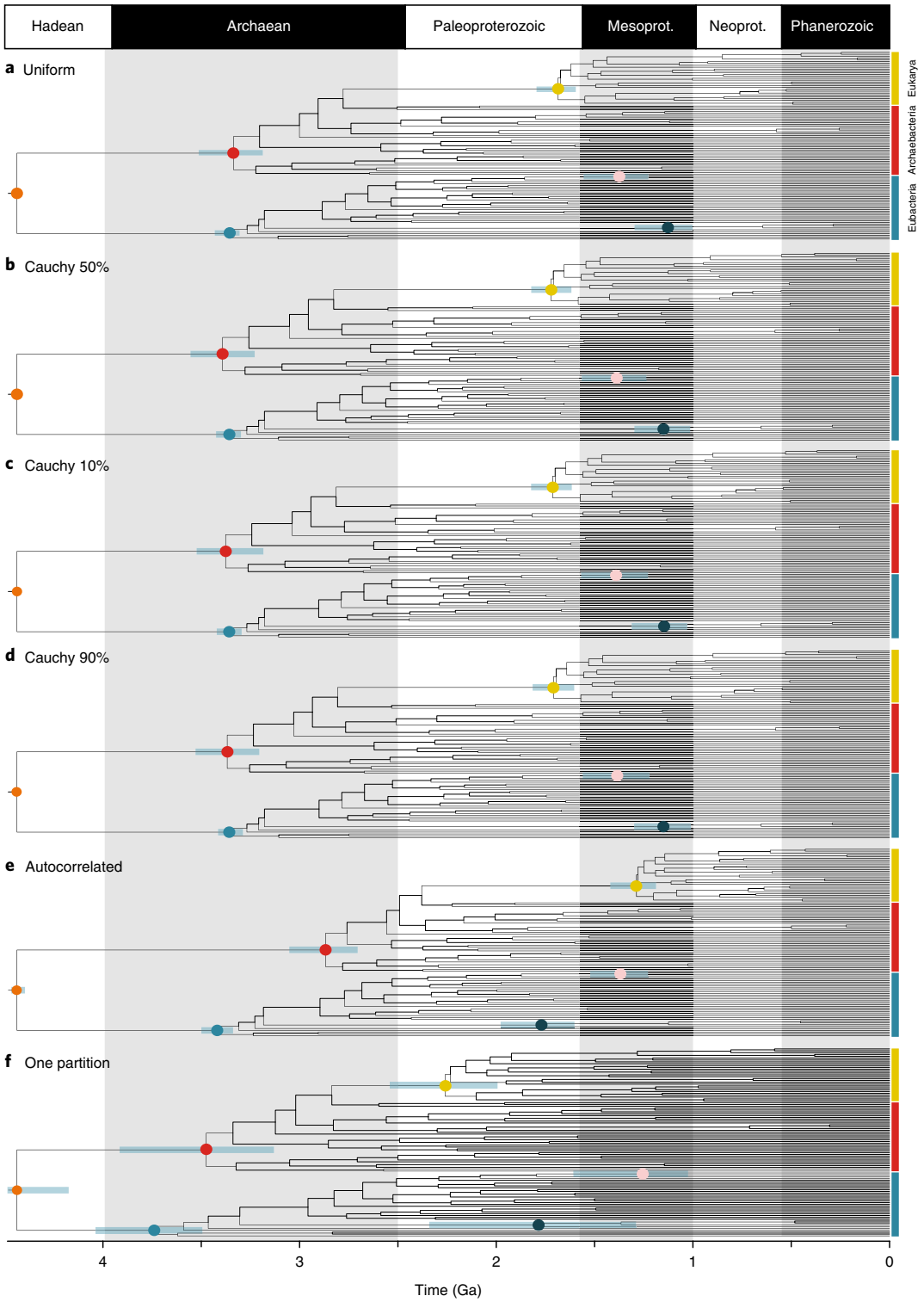

Fig. 1 | Posterior time estimates under different parameters. a, Posterior time estimates when using a uniform calibration density prior distribution, reflecting a lack of information about the divergence time relative to the fossil constraint. **b**, Cauchy 50% maximum calibration density prior distribution, reflecting a view that the divergence date should fall between the constraints. **c**, Cauchy 10% maximum calibration density prior distribution, reflecting a view that the fossil prior is a good approximation of the divergence date. **d**, Cauchy 90% maximum calibration density prior distribution, reflecting a view that the fossil prior is a poor approximation of the divergence date, all with an uncorrelated clock model. **e**,**f**, Posterior age estimates when using a Cauchy 50% maximum calibration density prior distribution with an autocorrelated clock model (**e**) and with an uncorrelated clock model and a single partition scheme (**f**). All molecular clock analyses converged well. The coloured dots highlight specific nodes, with their respective confidence intervals displayed light blue bars (orange, LUCA; red, crown Archaeabacteria; blue, crown Eubacteria; yellow, crown Eukaryota; pink, alphaproterobacteria; dark blue, cyanobacteria). This figure illustrates how divergence times change as alternative approaches to modelling calibrations and the process of molecular evolution are implemented. Divergence estimates from **f** and their credibility intervals could be rejected based on an AIC test. The other results (**a**–**e**) cannot be rejected. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

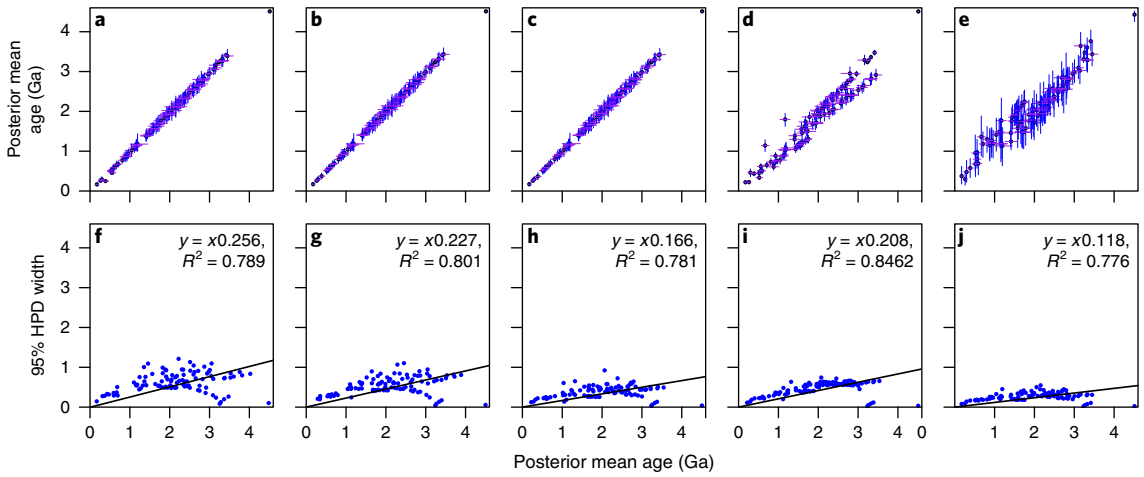

Fig. 2 | Changes in divergence times (Ga) that result from applying alternative parameters. a, Cauchy 50% maximum calibration density prior distribution versus uniform calibration density prior distribution. **b**, Cauchy 50% maximum calibration density prior distribution versus Cauchy 10% maximum calibration density prior distribution. **c**, Cauchy 50% maximum calibration density prior distribution versus Cauchy 90% maximum calibration density prior distribution. **d**, Cauchy 50% maximum calibration density prior distribution uncorrelated clock model versus Cauchy 50% maximum calibration density prior distribution autocorrelated clock model. **e**, Cauchy 50% maximum calibration density prior distribution in both cases for the 29-partition scheme versus the 1-partition scheme. **f**–**j**, Results of adding additional genes as infinite sites plots: 5-gene dataset (**f**); 10-gene dataset (**g**); 15-gene dataset (**h**); 20-gene dataset (**i**); 29-gene dataset (**j**). Blue dots denote node dates. HPD, highest posterior density.

conserved, mainly ribosomal, universally distributed proteins (see Supplementary Information) using a relaxed molecular clock modelled in a Bayesian framework.

Results

Analytical choices can deeply affect molecular clock posterior age estimates[32] and we explored a range of prior probability distributions to model our fossil calibrations and estimate conservative credibility intervals for our divergence times. Initially, we applied a hard maximum of 4.52 Ga (the age of the Moon-forming impact) to the root of our tree and used uniform age priors (reflecting agnosticism about divergence timing relative to constraints) to the other fossil calibrations (Fig. 1a). These analyses assumed an uncorrelated molecular clock model and produced the amino acid substitution processes using optimal gene-specific substitution models. Subsequently, we explored the impact of using calibration protocols based on non-uniform age priors. First, we implemented a truncated Cauchy distribution with the mode located halfway between the minimum and maximum bounds, reflecting a prior view that true divergence times should fall between the minimum and maximum calibration points (Fig. 1b). In two subsequent analyses we applied a skewed Cauchy distribution such that the mode shifted towards the minimum or the maximum constraint, reflecting prior views that the fossils used to calibrate the tree are either very good (Fig. 1c) or very poor (Fig. 1d) proxies of the true divergence times. Our results proved robust to the use of different calibration strategies, only identifying some variability in the size of the recovered credibility intervals (Fig. 2a–c).

We explored the impact of different strategies for modelling both the molecular clock (Fig. 1e) and the amino acid substitution process (Fig. 1f). Only minimal differences in posterior ages were found between analyses using an uncorrelated or autocorrelated clock (Fig. 2d). Consistently, Bayesian cross-validation indicated that the two models do not differ significantly in their fit to the data (cross-validation score = 0.7 ± 2.96816 in favour of the uncorrelated clock). In contrast, using a single substitution model across the 29 genes or using an optimal set of gene-specific substitution models inferred using PartitionFinder[33] resulted in very different age estimates (Figs. 1f and 2e). Using a single substitution model recovered larger credibility intervals (Fig. 2e) with a more homogeneous distribution of branch lengths across the tree, and older divergence times (compare Fig. 1f and Fig. 1a–d). An Akaike information criterion (AIC) test indicated that the partitioned model provides a significantly better fit to the data (AIC score = 565.21 in favour of 29 gene-specific models), allowing the rejection of the divergence times obtained with a single substitution model. As expected, divergence times estimated from individual genes were much less precise, although posterior age estimates overlap well (Supplementary Section 4.1). This indicates that the genes comprising our dataset encode a congruent signal and the timescale inferred from the combined analysis is not biased by single gene outliers. Furthermore, their combination improves the precision of the clade age estimates (Fig. 2f–j), which are clearly informed by the data (Supplementary Section 4.2). We tested the effect of taxonomic sampling by doubling the number of cyanobacteria and alphaproteobacteria in our dataset. We then explored the effect of phylogenetic uncertainty by dating a tree compatible with Woese's three-domains hypothesis[34] and by dating all 15 trees in the 95% credible set of trees from our phylogenetic analysis (Supplementary Sections 4.3 and 4.4). Further analyses that used co-estimation of tree and topology (Supplementary Section 4.5)[35] did not reach convergence (Supplementary Section 4.6), but the results recovered were congruent with those obtained from well-converged analyses (Supplementary Section 4.4) where topology and time were inferred sequentially (see the caption of Supplementary Section 4.5 for a discussion). Overall, the outcome of these experiments demonstrates that our original results are robust to topological uncertainty and the use of differential taxonomic sampling (Supplementary Sections 4.3–4.5).

It is not possible to discriminate between the competing calibration strategies that reflect different interpretations of the fossil record. Similarly, our model selection test indicated that the autocorrelated and independent-rates clock models fit the data equally well. Thus, in establishing an accurate timescale of life, we integrated over the uncertainties associated with the results from all these analyses (Fig. 3). The joint 95% credibility intervals reject a post-late heavy bombardment (~3,900 million years ago (Ma)[36] emergence of LUCA (4,519–4,477 Ma). The crown clades of the primary divisions of life, Archaebacteria and Eubacteria emerged over one billion years after LUCA in the Mesoarchaean–Neoarchaean. The earliest conclusive evidence of cellular life (Strelley Pool Formation,

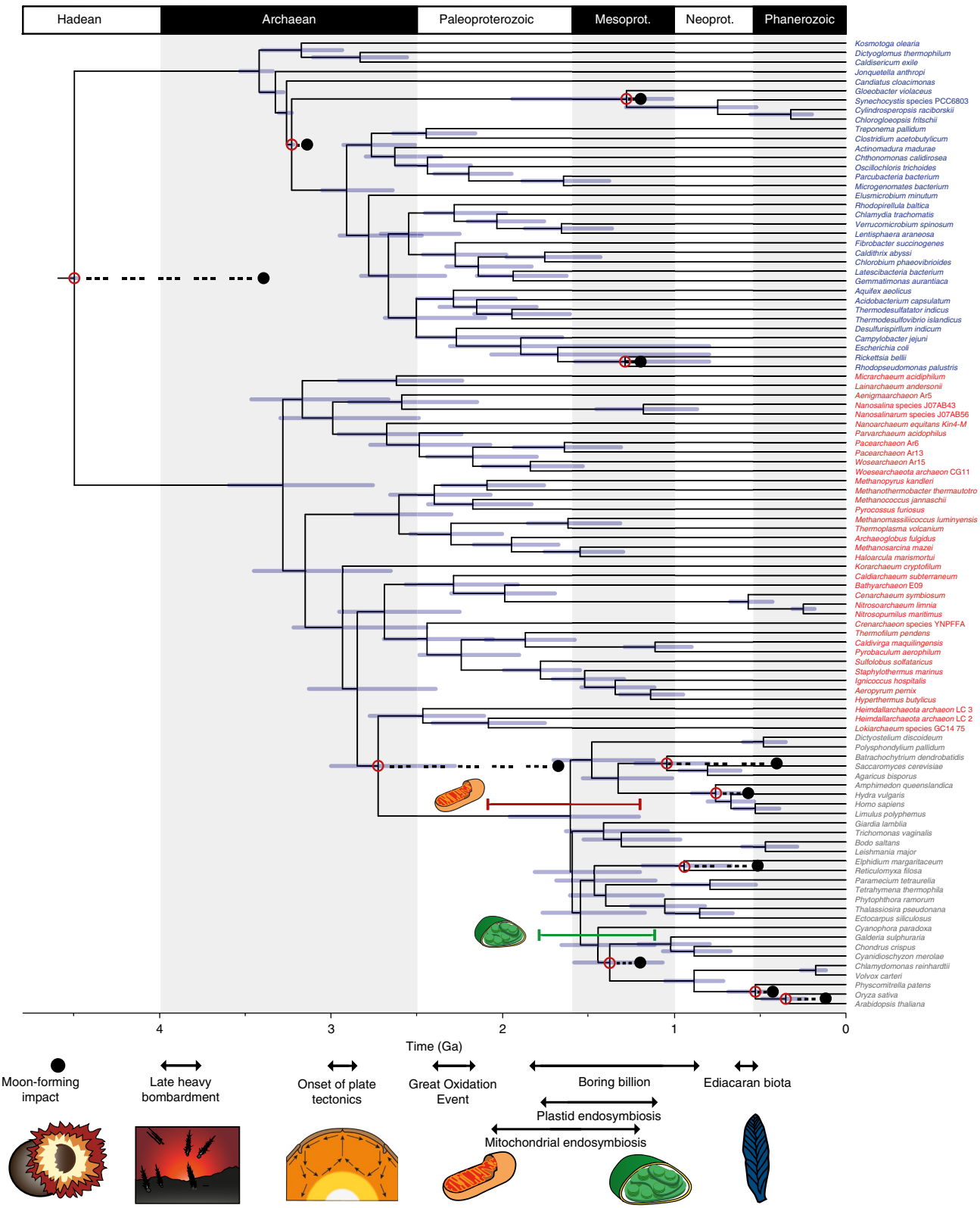

Fig. 3 | A tree combining uncertainties from approaches using uncorrelated and autocorrelated clock models and different calibration density distributions. Tip labels are shown for Eukaryota (grey), Archaeabacteria (red) and Eubacteria (blue). The purple bars denote the credible intervals for each node. Red dots highlight calibrated nodes, and corresponding black dots highlight the age of the minimum bound of its corresponding calibration. The phylogenetic relationships of the mitochondrion within Alphaproteobacteria are still debated[56,74–76], and it is unclear whether the free-living ancestor of the mitochondrion was a crown or stem representative of this group. The red bar above the crown eukaryote node denotes the time period during which the mitochondrial endosymbiosis may have occurred. The green bar denotes the time during which the plastid endosymbiosis may have occurred. Important events in Earth and life history are indicated along the base of the figure. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

Australia[2]) falls within the 95% credibility intervals for the ages of the last common ancestors of both clades, indicating that these fossils might belong to one of the two living prokaryotic lineages.

Discussion

Methanogenesis is classically associated with Euryarchaeota. Our estimate for the age of crown Euryarchaeota (2,881–2,425 Ma) is consistent with carbon isotope excursions indicating the presence of methanogens by 2 Ga[37], but is substantially younger than the earliest possible evidence of biogenic methane in the geochemical record at ~3.5 Ga[38,39]. If the geochemical evidence is correct, our timescale implies that methanogenesis predated the origin of Euryarchaeota. This hypothesis would be consistent with recent environmental genomic surveys indicating that other archaeal lineages may also be capable of methane metabolism[40] or methanogenesis[41], and that metabolisms using the Wood–Ljungdahl pathway to fix carbon minimally evolved in stem archaebacteria[42,43] and might have been a characteristic of LUCA[43–45].

The Great Oxidation Event (GOE; ~2.4 Ga) was perhaps the most significant episode in the Proterozoic[46], fundamentally changing the chemistry of Earth's atmosphere and oceans, and probably altering temperature. It has been causally associated with the evolution of Cyanobacteria, as a consequence of their oxygen release[28,47], and implicated as an extrinsic driver of eukaryotic evolution[48]. Our timescale indicates that crown Cyanobacteria and crown Eukaryota significantly postdate the GOE. Crown Cyanobacteria diverged 1,947–1,023 Ma, precluding the possibility that oxygenic photosynthesis emerged in the cyanobacterial crown ancestor. However, the Cyanobacteria separated from other eubacterial lineages (Fig. 3), including the non-photosynthetic sister group of the Cyanobacteria (Melanibacteria; Supplementary Section 4.3) in the Archaean, before the GOE, consistent with the view that oxygenic photosynthesis evolved along the cyanobacterial stem[49], and compatible with a causal role of the total-group Cyanobacteria in the GOE.

Crown Eukaryota diverged considerably after both the Eukaryota–Asgardarchaeota split and the GOE, in the middle Proterozoic (1,842–1,210 Ma). Our study strongly rejects the idea that eukaryotes might be as old as, or older than, prokaryotes[50], and agrees with a number of other studies that date the last eukaryote common ancestor (LECA) to the Proterozoic (~1,866–1,679 Ma)[51–53]. Within eukaryotes, the main extant clades emerged by the middle Proterozoic, including Opisthokonta (~1,707–1,125 Ma), Archaeplastida (~1,667–1,118 Ma) and SAR (stramenopiles (heterokonts), alveolates and Rhizaria; ~1,645–1,115 Ma). The symbiotic origin of the plastid occurred among stem archaeplastids (~1,774–1,118 Ma), and our 95% credibility interval for the origin of the plastid overlap with the results of other recent studies[28,50,54]. The relatively long stem lineage subtending LECA is intriguing. It is found using both uncorrelated and autocorrelated clock models (Figs. 1e and 2d), and disappears only if a poorly fitting single substitution model is used (Figs. 1f and 2e), suggesting that it is not a modelling artefact. Analyses excluding the hitherto unknown immediate living relatives of Eukaryota[9,31], Asgardarchaeota, had no significant impact on the span of the eukaryote stem lineage, suggesting that its length is robust to taxon sampling (Supplementary Section 4.7).

Our timescale for eukaryogenesis rejects the hypothesis of an inextricable link between the GOE and the origin of eukaryotes[48]. Competing hypotheses for eukaryogenesis hinge on the early versus late acquisition of mitochondria relative to other key eukaryote characters[55–59]. Absolute divergence times cannot discriminate between these hypotheses. However, as the only proposed evidence in support of the mitochondria late[57] hypothesis have been shown to be artefactual[58], the similar age estimates for Alphaproteobacteria and LECA at this stage are most conservatively interpreted as indicating that the process of mitochondrial symbiosis underpinned a rapid process of eukaryogenesis. This process involved a large transfer of genes from the genome of the alphaproteobacterial symbiont to that of the archaeal host[59,60], as predicated on metabolism[55,61].

The search for the earliest fossil evidence of life on Earth has created more heat than light. Although the fossil record remains integral to establishing a timescale for the Tree of Life, it is not sufficient in and of itself. Our integrative molecular timescale encompasses the uncertainty associated with fossil, geological and molecular evidence, as well its modelling, allowing it to serve as a solid foundation for testing evolutionary hypotheses in deep time for clades that do not have a credible fossil record.

Methods

Dataset collation and phylogenetic analysis. The dataset consists of 102 species and 29 universally distributed, protein-coding genes (see Supplementary Information). All our data and scripts are available at https://bitbucket.org/bzxdp/betts_et_al_2017. Proteomes were downloaded from GenBank[62] and putative orthologues were identified using BLAST[63]. The top hits were compiled and aligned into gene-specific files in MUSCLE[64] and trimmed to remove poorly aligned sites using Trimal[65]. To minimize the possible inclusion of paralogues and laterally transferred genes, we generated gene trees (under CAT-GTR + G) in PhyloBayes[66] and excluded sequences when the tree topology suggested that they might have been paralogues. The sequences were then concatenated into a supermatrix using FASconCAT[67], and phylogenetic analyses were performed using PhyloBayes[66]. The superalignment was initially analysed under both GTR + G and CAT-GTR + G[68]. RogueNaRok[69] was used to identify rogue taxa, and analyses were repeated (under both GTR + G and CAT-GTR + G) after unstable taxa were excluded. One final analysis was performed that included only the eukaryotic sequences in our dataset (under CAT-GTR + G). For all PhyloBayes analyses, convergence was tested in PhyloBayes using BPCOMP and TRACECOMP.

Calibrations. In total, we used 11 calibrations spread throughout the tree but mainly found within the Eukaryotes as this group has the best fossil record. Calibration choice was carried out conservatively using coherent criteria[4]. Full details of each calibration used can be found in the Supplementary Information.

MCMCTree analysis. For our clock analyses, we used a constraint tree based on our CAT-GTR + G and GTR + G trees (Supplementary Sections 3.2, 3.3 and 4; see the results of phylogenetic analyses in the Supplementary Information for details). The complete phylogeny was rooted to separate Eubacteria from the other lineages (that is, Archaeabacteria and Eukayota). To select the amino acid model to be used in our molecular clock analyses, we used PartitionFinder version 1.1.1 (ref. [33]). Divergence time estimation was carried out using the approximate likelihood calculation in MCMCTree version 4.9 (ref. [70]). We set four different calibration density distributions: uniform, skewed towards the minimum, skewed towards the maximum and midway between these two dates. For this, we used the Uniform and Cauchy models within MCMCTree, which can be set to place the maximum probability of the node falling in a certain space between the calibrations. The values for these were first produced using MCMCTreeR (https://github.com/PuttickMacroevolution/MCMCTreeR) code in R[71]. We investigated two strategies to model amino acid sequence evolution: a single WAG + G model or the optimal partitioned model suggested by PartitionFinder. The optimal partitioned model used 29 gene-specific models (28 LG + G and one WAG + G). The AIC was used to test whether using a single model or a partitioned model provided a better fit to the data. Rate variation across lineages was modelled using both an autocorrelated and uncorrelated clock model. Bayesian cross-validation was used to test whether one of the two considered, relaxed molecular clock models best fitted the data (implemented in PhyloBayes).

In all our molecular clock analyses, we applied a soft tail of 2.5% to the upper calibration bound and a hard minimum, apart from the root node (to which a hard maximum was applied) and the nodes calibrated using *Bangiomorpha*[72] (to which a soft minimum tail of 2.5% was applied). For all molecular clock analyses, convergence was tested in Tracer[73] by comparing plots of estimates from the two independent chains and evaluating whether—for each model parameter and divergence time estimate—the effective sample size was sufficiently large. All reported molecular clock analyses reached excellent levels of convergence.

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

Data availability. All data that support the findings of this study are available from Bitbucket: https://bitbucket.org/bzxdp/betts_et_al_2017.

Received: 15 April 2018; Accepted: 13 July 2018;
Published online: 20 August 2018

References

1. Dos Reis, M., Donoghue, P. C. J. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71–80 (2016).
2. Wacey, D. *Early Life on Earth: a Practical Guide* Vol. 31 (Springer, New York, 2009).
3. Inoue, J., Donoghue, P. C. J. & Yang, Z. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74–89 (2009).
4. Parham, J. F. et al. Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359 (2012).
5. Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R. & Donoghue, P. C. J. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc. R. Soc. B* **282**, 20141013 (2014).
6. Davín, A. A. et al. Gene transfers can date the tree of life. *Nat. Ecol. Evol.* **2**, 904–909 (2018).
7. Lozano-Fernandez, J., dos Reis, M., Donoghue, P. C. J. & Pisani, D. RelTime rates collapse to a strict clock when estimating the timeline of animal diversification. *Genome Biol. Evol.* **9**, 1320–1328 (2017).
8. Pisani, D. & Liu, A. G. Animal evolution: only rocks can set the clock. *Curr. Biol.* **25**, R1079–R1081 (2015).
9. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
10. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).
11. Dodd, M. S. et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* **543**, 60–64 (2017).
12. Pflug, H. D. & Jaeschke-Boyer, H. Combined structural and chemical analysis of 3,800-Myr-old microfossils. *Nature* **280**, 483–486 (1979).
13. Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J. & Chivas, A. R. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* **537**, 535–538 (2016).
14. Rosing, M. T. ^{13}C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
15. Mojzsis, S. J.et al. Evidence for life on Earth before 3,800 million years ago. *Nature* **384,** 55–59 1996).
16. Van Zuilen, M. A., Lepland, A. & Arrhenius, G. Reassessing the evidence for the earliest traces of life. *Nature* **418**, 627–630 (2002).
17. Horita, J. & Berndt, M. E. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285**, 1055–1057 (1999).
18. Lepland, A., Arrhenius, G. & Cornell, D. Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Res.* **118**, 221–241 (2002).
19. Sugitani, K. et al. Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology* **13**, 507–521 (2015).
20. Sugitani, K. et al. Biogenicity of morphologically diverse carbonaceous microstructures from the ca. 3400 Ma Strelley Pool Formation, in the Pilbara Craton, Western Australia. *Astrobiology* **10**, 899–920 (2010).
21. Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K. & Takeuchi, M. Microfossil assemblage from the 3400 Ma Strelley Pool Formation in the Pilbara Craton, Western Australia: results form a new locality. *Precambrian Res.* **226,** 59–74 2013).
22. Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4**, 698–702 (2011).
23. Lepot, K. et al. Texture-specific isotopic compositions in 3.4 Gyr old organic matter support selective preservation in cell-like structures. *Geochim. Cosmochim. Acta* **112**, 66–86 (2013).
24. Wacey, D., McLoughlin, N., Whitehouse, M. J. & Kilburn, M. R. Two coexisting sulfur metabolisms in a ca. 3400 Ma sandstone. *Geology* **38**, 1115–1118 (2010).
25. Wacey, D. Stromatolites in the ~3400 Ma Strelley Pool Formation, Western Australia: examining biogenicity from the macro-to the nano-scale. *Astrobiology* **10**, 381–395 (2010).
26. Abramov, O. & Mojzsis, S. J. Microbial habitability of the Hadean Earth during the late heavy bombardment. *Nature* **459**, 419–422 (2009).
27. Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).
28. Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl Acad. Sci. USA* **114**, E7737–E7745 (2017).
29. Bengston, S. et al. Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLoS Biol.* **15**, e2000735 (2017).
30. Lamb, D. M., Awramik, S. M., Chapman, D. J. & Zhu, S. Evidence for eukaryotic diversification in the ~1800 million-year-old Changzhougou Formation, North China. *Precambrian Res.* **173**, 93–104 (2009).
31. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
32. Warnock, R. C. M., Yang, Z. & Donoghue, P. C. J. Exploring uncertainty in the calibration of the molecular clock. *Biol. Lett.* **8**, 156–159 (2012).
33. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
34. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
35. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
36. Chapman, C. R., Cohen, B. A. & Grinspoon, D. H. What are the real constraints on the existence and magnitude of the late heavy bombardment? *Icarus* **189**, 233–245 (2007).
37. Hayes, J. M. in *Early life on Earth* Vol. 84, 220–236 (Columbia University Press, New York, 1994).
38. 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).
39. Wolfe, J. & Fournier, G. P. Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol.* **2**, 897–903 (2018).
40. Evans, P. N. et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
41. Vanwonterghem, I. et al. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 16170 (2016).
42. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl Acad. Sci. USA* **114**, 4602–4611 (2017).
43. Weiss, M. C. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
44. Sousa, F. L., Nelson-Sathi, S. & Martin, W. F. One step beyond a ribosome: the ancient anaerobic core. *Biochim. Biophys. Acta* **1857**, 1027–1038 (2016).
45. Borrel, G., Adam, P. S. & Gribaldo, S. Methanogenesis and the Wood–Ljungdahl pathway: an ancient, versatile, and fragile association. *Genome Biol. Evol.* **8**, 1706–1711 (2016).
46. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
47. Schirrmeister, B. E., de Vos, J. M., Antonelli, A. & Bagheri, H. C. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc. Natl Acad. Sci. USA* **110**, 1791–1796 (2013).
48. Knoll, A. H. & Nowak, M. A. The timetable of evolution. *Sci. Adv.* **3**, e1603076 (2017).
49. 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).
50. Kurland, C. G., Collins, L. J. & Penny, D. Genomics and the irreducible nature of eukaryote cells. *Science* **312**, 1011–1014 (2006).
51. Chernikova, D., Motamedi, S., Csürös, M., Koonin, E. V. & Rogozin, I. B. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6**, 26 (2011).
52. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *CSH Perspect. Biol.* **6**, a016139 (2014).
53. 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. USA* **108**, 13624–13629 (2011).
54. Shih, P. M. & Matzke, N. J. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl Acad. Sci. USA* **110**, 12355–12360 (2013).
55. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **12**, 449–455 (2014).
56. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
57. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
58. Martin, W. F. et al. Late mitochondrial origin is an artifact. *Genome Biol. Evol.* **9**, 373–379 (2017).
59. Pisani, D., Cotton, J. A. & McInerney, J. O. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760 (2007).
60. Ku, C. et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).
61. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).

62. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
65. 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).
66. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
67. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
68. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
69. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
70. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
71. R Core Development Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
72. Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* **250**, 104–108 (1990).
73. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* https://doi.org/10.1093/sysbio/syy032 (2018).
74. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
75. Esser, C. et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643–1660 (2004).
76. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome phylogenies indicate a meaningful α-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.* **23**, 74–85 (2006).

Acknowledgements

H.C.B. was supported by a NERC GW4 PhD studentship. J.W.C. was supported by a BBSRC SWBio PhD studentship. M.N.P. was supported by an 1851 Royal Commission Fellowship. P.C.J.D. was supported by BBSRC grant BB/N000919/1. T.A.W. is supported by a Royal Society Fellowship and NERC grant NE/P00251X/1.

Author contributions

D.P., P.C.J.D. and T.A.W. designed the study. H.C.B. assembled the datasets and performed the phylogenetic and molecular clock analyses. M.N.P. and J.W.C. contributed further molecular clock analyses. H.C.B., D.P., P.C.J.D. and T.A.W. wrote the manuscript. All authors edited the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0644-x.

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

Correspondence and requests for materials should be addressed to D.P.

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

nature research

Corresponding author(s): Davide Pisani

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 Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☒	☐	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All used data was obtained from the NCBI website and is publicly available.
Data analysis	Muscle (Edgar (2004) NAR), was used to align the sequences. TrimAL (Capella-Gutierrez et al. (2009) Bioinformatics) was used to remove poorly aligned sites. FasConcat (Kuck and Meusemann 2010 Mol Phylogenet Evol) was used to concatenate single gene alignments into our 29 gene superalignment. RogueNaRok (Aberer et al. (2013) Systematic Biology) was used to identify rogue taxa. Phylobayes MPI version 1.7a (Lartillot et al. 2009 Bioinformatics) was used for all Bayesian phylogenetic analyses and to compare alternative molecular clock models using 10-fold Bayesian Crossvalidation. PartitionFinder (Lanfear 2012 Mol Biol Evol) was used to estimate the best fitting models for individual genes that we used for our molecular clock analyses. PAML 4.9 (Yang 2007 Mol Biol Evol) was used for all molecular clock analyses. MCMCTREER. We also used a bespoke software written by Mark Puttick (one of the co-authors). The software estimates the parameters

for the Cauchy distributions to be used in MCMCTREE to define densities representing fossil calibrations.
MCMCTREER is available in GitHub and we provide a link in the paper (https://github.com/PuttickMacroevolution/MCMCTreeR).

MrBayes was used to carry out co-estimation of time and topology (mrbayes.sourceforge.net/).

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/reviewers upon request. 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

Accession numbers for all sequences in our study are reported in supplementary information. All our multiple sequence alignments have been deposited in a public data repository and are freely and publicly available https://bitbucket.org/bzxdp/betts_et_al_2017.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

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

Study description	Our study present a phylogenomic analysis and a large scale molecular divergence time analysis to date the history of life on Earth and an associated reassessment of the vailidity of the fossil record of early life, based on published information and publicly available data.
Research sample	Sample size is important in phylogenomics but it is not defined as in standard statistical analyses. Our molecular dataset includes 29 genes, these are all the genes we could identify that are shared across all lineages of life and do not include paralogs and xenologs – explained in the paper. In total the 29 genes correspond to an alignment of 14,645 amino acid positions.
Sampling strategy	When defining a dataset for phylogenetic/molecular clock analyses it is fundamental to include all species of interest, while maintaining a balanced taxon sampling. Our dataset included 102 species of which 29 eukaryotes, 35 eubacteria and 38 archaebacteria. Our dataset is thus well balanced, there are about the same number of species for each lineage, and it covers the necessary taxonomic diversity.
Data collection	Molecular data was obtained from NCBI (all publicly available). Fossil information was obtained from literature searches. All analyses were carried out by Holly Betts.
Timing and spatial scale	This does not really apply to our type of data (I think). But all data were collected from papers and online repositories prior to September the 1st 2017
Data exclusions	As it is standard in phylogenomics and molecular clock analyses some data were excluded. For both phylogenetic reconstruction and molecular dating we excluded poorly aligned sites using a well-established standard bioinformatic tool – TrimAl, Capella-Gutierrez et al. (2009) Bioinformatics. In addition, for the phylogenetic analyses we investigated the impact of "rogue taxa". These are taxa that are phylogenetically unstable, depress support values and can cause Bayesian analyses to fail to reach convergence (see Pisani et al. 2015 PNAS for a recent example). We identified 5 unstable taxa that were excluded in some phylogenetic analyses. Unstable taxa were identified using well-established software – RogueNaRok – Aberer et al. (2013) Systematic Biology. Calibrations: A large number of putative fossils are constantly being described by palaeontologists. However, most of these fossils cannot be used for calibrating nodes in molecular clock analyses. There are many reasons why this happens, for example, a specific formation might not be dated precisely enough, or a fossil might lack the specific characters that are needed to certify its biogenic origin. This is a particularly serious problem with the fossil record of early life. We reviewed the fossil record of early life in detail and excluded all the fossils that did not meet the criteria necessary to define a good quality calibration. To reach this aim we followed well-established criteria (Parham et al. 2011 Systematic Biology). All the above methods are clearly described in the paper
Reproducibility	All findings in the published paper are based on converged Bayesian analyses. This is tested by running analyses independently multiple time and implies that the results are reproducible by default.

Randomization	Does not apply to our study
Blinding	Blinding was not relevant to this study. No blinding is done in phylogenomics and comparative genomics more broadly.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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