



Betts, H. C., Puttick, M. N., Clark, J. W., Williams, T. A., Donoghue, P. C. J., & Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology and Evolution*, 2(10), 1556-1562. <https://doi.org/10.1038/s41559-018-0644-x>

Peer reviewed version

Link to published version (if available):
[10.1038/s41559-018-0644-x](https://doi.org/10.1038/s41559-018-0644-x)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Springer Nature at <https://www.nature.com/articles/s41559-018-0644-x#Abs1>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

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

Holly C. Betts¹, Mark N. Puttick¹, James W. Clark¹, Tom A. Williams^{1,2}, Philip C. J. Donoghue¹, Davide Pisani^{*1,2}

Schools of Earth¹ and Biological² Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

* Author for correspondence Email: davide.pisani@bristol.ac.uk

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 that the last universal common ancestor of cellular life (LUCA) predated the end of late heavy bombardment (>3.9 Ga). The crown clades of the two primary divisions of life, Eubacteria and Archaeobacteria, 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 photosynthesis evolved within the cyanobacterial stem-lineage. Modern eukaryotes emerged late in Earth history (<1.84 Ga), falsifying the hypothesis that the GOE facilitated their radiation. The symbiotic origin of mitochondria, at 2.053 – 1.21 Ga reflects a late origin of the eukaryotes, that do not constitute a primary lineage of life.

Examining the emergence of Life and its subsequent evolution has traditionally been carried out via interpretation of the fossil record. However, this record, especially when looking at the earliest scions of life, is minimal and made harder to interpret due to difficulties substantiating relationships within the earliest branching lineages of the tree of life^{1,2}. Despite being problematic the fossil record has been the main source of information informing the timeline of life's evolution. We attempt to shed light on this early period by presenting a molecular timescale utilising 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, e.g. horizontal gene transfers⁶. 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 practice⁴ for the fundamental clades within the Tree of Life, drawing on multiple lines of evidence, including physical fossils, biomarkers and isotope geochemistry². 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 Eoarchaeon, including microfossils^{11,12}, stromatolites¹³ and isotope data^{14,15} from the ~3.8 Ga Isua Greenstone Belt (Greenland). However, these records have been contested¹⁶⁻¹⁸. Microfossils from the ~3.4 Ga Strelley Pool Formation, Australia, are the oldest conclusive evidence to constrain the age of LUCA¹⁹. The fossils, many of which are arranged in chains of cells, have been demonstrated 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 pseudofossils². The Strelley Pool Formation also

contains other microfossils²⁰⁻²², in association with both distinct $\delta^{13}\text{C}_{\text{org}}$ and $\delta^{13}\text{C}_{\text{inorg}}$ ²³ and pyrite indicative of sulphur metabolisms²⁴, along with stromatolites that exhibit biological structure²⁵. 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 sterilization²⁶. However, the last formative stage of Earth's formation, the Moon-forming impact, melted and sterilised the planet. The oldest fossil remains that can be ascribed to crown-Eukaryota is ~ 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 Chitrukoot Formation have been tentatively interpreted as red algae²⁹, 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 Eukaryota are acritarchs from the >1.6191 Ga Changcheng Formation, North China³⁰, 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 to any specific crown-eukaryote clade, only allowing us to use these records to constrain minimally the timing of divergence between the Eukaryota and their archaeobacterial 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, i.e. the Moon-forming impact. These key time constraints were combined with nine others (See SI) to calibrate a timescale of life estimated from a dataset of 29 highly conserved, mainly ribosomal, universally distributed proteins (see SI) using a relaxed molecular clock modelled in a Bayesian framework.

Results

Analytical choices can deeply affect molecular clock posterior age estimates³² 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. Firstly, 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 calibration points (Fig. 1b). We then skewed the 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 an 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 PartitionFinder³³ resulted in very different age estimates (Fig. 1f, 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 (SI S4.1). 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 (SI S4.2). 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 hypothesis³⁴ and by dating all the 15 trees in the 95% credible set of trees from our phylogenetic analysis (S4.3, S4.4). Further analyses that used co-estimation of tree and topology (S4.5)³⁵ did not reach convergence (S4.6), but recovered results that are congruent with those obtained from well converged analyses (reported in S.4.4) where topology and time were sequentially inferred – see S4.5 for a discussion. Overall, it appears that our results are robust to topological uncertainty and the use of differential taxonomic sampling (SI S4.3 – S4.5).

It is not possible to discriminate between the competing calibration strategies which reflect different interpretations of the fossil record, similarly, our model selection test indicated that the autocorrelated and the independent-rates clock models fits the data equally well. Thus, in establishing an accurate timescale of life we integrated over the uncertainties associated with results from all these analyses (Fig. 3). The joint 95% credibility intervals reject a post late heavy bombardment (~3900 Ma)³⁶ emergence of LUCA (4519-4477 Ma). The crown clades of the primary divisions of life, Archaeobacteria and Eubacteria, emerged over a billion years after LUCA in the Mesoarchaeoan-Neoarchaeoan. The earliest conclusive evidence of cellular life (Strelley Pool Formation, Australia²) 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 (2881-2425 Ma) is consistent with carbon isotope excursions indicating the presence of methanogens by 2 Ga³⁷ 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 pre-dated 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 metabolism⁴⁰ or methanogenesis⁴¹, and that metabolisms using the Wood-Ljungdahl pathway to fix carbon minimally evolved in stem-archaebacteria^{42(WAS6),43} and might have been a characteristic of LUCA⁴³⁻⁴⁵.

The GOE (~2.4 Ga) was perhaps the most significant episode in the Proterozoic⁴⁶, fundamentally changing the chemistry of Earth's atmosphere, oceans, and likely altering temperature. It has been causally associated with the evolution of cyanobacteria, as a consequence of their oxygen release^{47,28}, and implicated as an extrinsic driver of eukaryotic evolution⁴⁸. Our timescale indicates that crown-Cyanobacteria and crown-Eukaryota significantly postdate the GOE. Crown-Cyanobacteria diverged 1947–1023 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; SI 4.3) in the Archaean, prior to the GOE, consistent with the view that oxygenic photosynthesis evolved along the cyanobacterial stem⁴⁹ 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 (1842-1210 Ma). Our study strongly rejects the idea that eukaryotes might be as old, or older than prokaryotes⁵⁰, and agrees with a number of other studies that dates the Last

Eukaryote Common Ancestor (LECA) to the Proterozoic (~1866-1679 Ma)⁵¹⁻⁵³. Within eukaryotes, the main extant clades emerged by the middle Proterozoic, including Opisthokonta (~1707-1125 Ma), Archaeplastida (~1667-1118 Ma) and SAR (~1645-1115 Ma). The symbiotic origin of the plastid occurred among stem archaeplastids (~1774-1118 Ma), and our 95% credibility interval for the origin of the plastid overlap with the results of other recent studies^{28,50,53}. The relatively long stem-lineage subtending LECA is intriguing. It is found using both uncorrelated and autocorrelated clock models (Fig. 1e, 2d), and disappears only if a poorly-fitting, single substitution model is used (Fig. 1f, 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 (SI 4.7).

Our timescale for eukaryogenesis rejects the hypothesis of an inextricable link between the GOE and the origin of eukaryotes⁴⁸. Competing hypothesis for eukaryogenesis hinge on the early versus late acquisition of mitochondria relative to other key eukaryote characters^{55,56-59}. Absolute divergence times cannot discriminate between these hypotheses. However, as the only proposed evidence in support of the mitochondria late⁵⁷ hypothesis have been shown to be artifactual⁵⁸, the similar age estimates for Alphaproteobacteria and LECA, at this stage are most conservatively interpreted as indicating that the process of mitochondrial symbiosis drove a rapid process of eukaryogenesis. This process involved 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. Though 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 conclusive fossil record.

Materials and Methods

Dataset collation and Phylogenetic analysis

The dataset consists of a 102 species and 29, universally distributed, protein-coding genes (see SI). All our data and scripts are available at https://bitbucket.org/bzxdp/betts_et_al_2017. Proteomes were downloaded from GenBank⁶² and putative orthologs were identified using BLAST⁶³. The top hits were compiled and aligned into gene specific files in MUSCLE⁶⁴ and trimmed to remove poorly aligned sites using Trimal⁶⁵. To minimise the possible inclusion of paralogs and laterally transferred genes, we generated gene trees (under CAT-GTR+G) in PhyloBayes⁶⁶ and excluded sequences when the tree topology suggested that they might have been paralogs. The sequences were then concatenated into a supermatrix using FASconCAT⁶⁷, and phylogenetic analyses were performed using PhyloBayes⁶⁶. The superalignment was initially analysed under both GTR+G and CAT-GTR+G⁶⁸. RogueNaRok⁶⁹ was used to identify rogue taxa, and analyses were repeated (under both GTR+G and CAT-GTR+G) after having excluded unstable taxa. 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 criteria⁷⁰. 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 (Fig. S3.2, S3.3 and S4) – see results of phylogenetic analyses in SI for details. The complete phylogeny was rooted to separate the Bacteria from the other lineages (i.e. Archaea and Eukayota). To select the amino acid model to be used in our molecular clock analyses we used PartitionFinder v1.1.1⁷¹. Divergence time estimation was carried out using the approximate likelihood calculation in MCMCTree v4.9⁷². We set 4 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 MCMCTreeR⁷³ code in R⁷⁴. We investigated two strategies to model amino acid sequence evolution: a single WAG+G model or the optimal partitioned model suggested by PartitionFinder. The latter used 29 gene-specific models (28 LG+G and one WAG+G). The Akaike Information Criterion 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 an 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 a soft tail of 2.5% was applied to the upper calibration bound and a hard minimum. This apart from the root node to which a hard maximum was applied, and the nodes calibrated using *Bangiomorpha*⁷⁵ to which a soft minimum tail of 2.5% was applied. For all Molecular clock analyses convergence was tested in Tracer⁷⁶ by comparing plots of estimates from the two independent chains and evaluating that, 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.

Data Availability. All our dataset are available in Bitbucket https://bitbucket.org/bzxdp/betts_et_al_2017.

- 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 Science & Business Media, 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. *Systematic Biology* **59**, 74-89 (2009).
- 4 Parham, J. F. *et al.* Best practices for justifying fossil calibrations. *Systematic Biology* **61**, 346–359 (2011).
- 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 Davine
- 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 Biology and Evolution* **9**, 1320-1328 (2017).
- 8 Pisani, D. & Liu, A. G. Animal evolution: only rocks can set the clock. *Current Biology* **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. ¹³C-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 (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 Research* **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 from a new locality. *Precambrian Research* **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. *Nature 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. *Geochimica et Cosmochimica 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**, (2000).
- 28 Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences* **114**, E7737-E7745, (2017).
- 29 Bengtson, S. *et al.*, Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLOS Biology*, **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 Research* **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. *Biology Letters* **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. *Molecular Biology and Evolution* **29**, 1695-1701 (2012).
- 34 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences* **74**, 5088-5090, (1977).
- 35 Drummond, A.J., Ho, S.Y.W, Phillips M.J. & Rambaut A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology* **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. Global methanotrophy at the Archean-Proterozoic transition. *Early life on Earth* **84**, 220-236 (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. M. & Fournier, G. P. Tunneling through time: Horizontal gene transfer constrains the timing of methanogen evolution. *BioRxiv* 129494 (2017).
- 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.* Methylophilic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature Microbiology* **1**, 16170 (2016).
- 42 Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences* **114**, 4602-4611 (2017).
- 43 Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. **1**, 16116 (2016).

- 44 Sousa, F. L., Nelson-Sathi, S. & Martin, W. F. One step beyond a ribosome: The ancient anaerobic core. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **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 Biology and Evolution* **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 Schirromeister, 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. *Proceedings of the National Academy of Sciences* **110**, 1791-1796 (2013).
- 48 Knoll, A. H. & Nowak, M. A. The timetable of evolution. *Science Advances* **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. *Biology 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. *Cold Spring Harb. 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. *Proceedings of the National Academy of Sciences* **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. *Proceedings of the National Academy of Sciences* **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. *Nature Reviews Microbiology* **12**, 449-455 (2014).
- 56 Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Current Biology* **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 Biology and Evolution* **9**, 373-379, (2017).
- 59 Pisani, D. *et al.* Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution* **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 Research* **41**, D36-D42 (2013).
- 63 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).
- 64 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **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. *Molecular Phylogenetics and Evolution* **56**, 1115-1118 (2010).
- 68 Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* **21**, 1095-1109, doi:10.1093/molbev/msh112 (2004).
- 69 Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology* **62**, 162-166 (2013).
- 70 Parham, J. F. *et al.* Best practices for justifying fossil calibrations. *Systematic Biology*, sy107 (2011).
- 71 Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution* **29**, 1695-1701 (2012).
- 72 Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
- 73 Puttick, M. N. MCMCTreeR. *GitHub repository*, (2017).
- 74 R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna, Austria* (2017).
- 75 Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* **250**, 104-108 (1990).
- 76 Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* **syy032**. (2018).
- 77 Martijn J., Vosseberg J., Guy L., Offre P., Ettema T.J.G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature*, **557**:101-105 (2018).
- 78 Esser C., Ahmadinejad N., Wiegand C., Rotte C., Sebastiani F., Gelius-Dietrich G., Henze K, Kretschmann E., Richly E., Leister D., Bryant D., Steel M.A., Lockhart P.J.,

Penny D., Martin W. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, **21**:1643-60 (2004).

- 79 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. *Molecular Biology and Evolution*, **23**:74–85 (2006).

Acknowledgments

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 a 1851 Royal Commission Fellowship, P.C.D. by a BBSRC grant BB/N000919/1. T.A.W. is supported by a Royal Society Fellowship and a NERC grant NE/P00251X/1.

Author contributions

D.P., P.C.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.D. and T.A.W. wrote the manuscripts. All authors edited the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Figure 1. Posterior time estimates when using (a) a Uniform calibration density prior distribution, reflecting a lack of information about the divergence time relative to the fossil constraint, (b) a Cauchy 50% maximum calibration density prior distribution, reflecting a view that the divergence date should fall between the constraints, (c) a Cauchy 10% maximum calibration density prior distribution, reflecting a view that the fossil prior is a good approximation of the divergence date and (d) a 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. Then posterior age estimates when using (e) a Cauchy 50% maximum calibration density prior distribution with an autocorrelated clock model and finally, (f) a Cauchy 50% maximum calibration density prior distribution, with an uncorrelated clock model and a single partition scheme. All molecular clock analysis converged well. The coloured dots highlight specific nodes with their respective confidence intervals displayed as bars in light blue; orange (LUCA), red (crown Archaeobacteria), blue (crown Eubacteria), yellow (crown Eukaryota), pink (alphaproterobacteria) and dark blue (cyanobacteria). This figure illustrates how divergence times change as alternative approaches to modelling calibrations and the process of molecular revolution change. Divergence estimates from Figure 1f and their credibility intervals could be rejected based on an AIC test. The other results (Figure 1a-e) cannot be rejected.

Figure 2. Changes in divergence times (billions of years before present) that result from applying alternative strategies, (a) Cauchy 50% maximum calibration density prior distribution vs Uniform calibration density prior distribution, (b) Cauchy 50% maximum calibration density prior distribution vs Cauchy 10% maximum calibration density prior distribution, (c) Cauchy 50% maximum calibration density prior distribution vs Cauchy 90% maximum calibration density prior distribution, (d) Cauchy 50% maximum calibration density prior distribution uncorrelated clock model vs Cauchy 50% maximum calibration density prior distribution autocorrelated clock model, (e) Cauchy 50% maximum calibration density prior distribution in both cases, 29 partition scheme vs 1 partition scheme. The lower row displays the results of adding additional genes as infinite sites plots, (f) 5 gene dataset, (g)

10 gene dataset, (h) 15 gene dataset, (i) 20 gene dataset and (j) 29 gene dataset. In all cases blue dots denote the node dates.

Figure 3. A tree combining uncertainties from approaches using uncorrelated and autocorrelated clock models and different calibration density distributions showing tip labels for Eukaryota (grey), Archaeabacteria (red), and Eubacteria (blue). The purple bars denote the credible intervals for each node. Red dots highlight calibrated nodes and the corresponding black dot highlights the age of the minimum bound of its corresponding calibration. The phylogenetic relationships of the mitochondrion within Alphaproteobacteria are still debated^{56,77-79}, and it is unclear whether the free-living ancestor of the mitochondrion was a crown or a stem representative of this group. The pink and red bars above the crown eukaryote node denote the time period in which crown alphaproteobacteria and stem alphaproteobacteria respectively were also present, and therefore in such time that the mitochondrial endosymbiosis may have occurred. The green bar lower down denotes the time in which the plastid endosymbiosis may have occurred. Some important events in the Earth's and life's history are indicated along the base of the figure.





