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