Estimating the timing of early eukaryotic diversification with multigene molecular clocks

Elucidating the Timing of Eukaryotic Diversification

Laura Wegener Parfrey[a,b,1,2], Daniel J.G. Lahr[a,b], Andrew H. Knoll[c,1], Laura A. Katz[a,b,1]

[a]Program in Organismic and Evolutionary Biology, University of Massachusetts, 611 North Pleasant Street, Amherst, MA 01003, USA

[b]Department of Biological Sciences, Smith College, 44 College Lane, Northampton, MA 01063, USA

[c]Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

Keywords: Molecular clock, microbial eukaryotes, Proterozoic oceans, microfossils

[1]To whom correspondence should be addressed. Email: Laura.Parfrey@Colorado.edu, Aknoll@oeb.harvard.edu, Lkatz@smith.edu

[2]Present address: Department of Chemistry and Biochemistry, University of Colorado, 215 UCB, Boulder, CO 80309, USA

Classification: Biological Sciences - Evolution

Although macroscopic plants, animals, and fungi are the most familiar eukaryotes, the bulk of eukaryotic diversity is microbial. Elucidating the timing of diversification among the more than 70 microbial lineages is thus key to understanding the evolution of eukaryotes. Here, we use taxon-rich multigene data combined with diverse fossils and a relaxed molecular clock framework to estimate the timing of the last common ancestor of extant eukaryotes and the divergence of major lineages. Overall, these analyses suggest that the last common ancestor lived between 1866 and 1679 million years ago (Ma), consistent with the earliest microfossils interpreted with confidence as eukaryotic. During this interval, the Earth's surface differed markedly from today; for example, the oceans were incompletely ventilated, with ferruginous and, after about 1800 Ma, sulfidic water masses commonly lying beneath moderately oxygenated surface waters. Our time estimates also indicate that the major clades of eukaryotes diverged before 1000 Ma, with most or all probably diverging before 1200 Ma. Fossils, however, suggest that diversity within major extant clades expanded later, beginning about 800 Ma, when the oceans began their transition to a more modern chemical state. Our molecular results are consistent with the geological record in terms of the timing of eukaryote origins. In combination, paleontological and molecular approaches indicate that long stems preceded diversification in the major eukaryotic lineages.

\body

Introduction

The antiquity of eukaryotes and the tempo of early eukaryotic diversification remain open questions in evolutionary biology. Proposed dates for the origin of the domain, based on the fossil record and molecular clock analyses, differ by up to two billion years (1). Putative biomarkers of early eukaryotes have been found in 2700 Ma rocks (2) and microfossils attributed to eukaryotes occur at about 1800 Ma (3). Such geological interpretations, which indicate a relatively early origin of nucleated cells, contrast with molecular clock studies that place the origin of eukaryotes at 1250–850 Ma (4, 5) and a controversial hypothesis that places eukaryogenesis at 850 Ma, rejecting both molecular clock estimates and the eukaryotic interpretation of all older fossils and biomarkers (6, 7).

Paleontologists generally agree that an unambiguous record of eukaryotic microfossils extends back to around 1800 Ma (3, 8, 9). Microfossils of this age are assigned to eukaryotes because they combine informative characters that include complex morphology (e.g., presence of processes and evidence for real-time modification of vegetative morphology), complex wall ultrastructure, and specific inferred behaviors (3, 9, 10). Despite being interpreted as eukaryotic, however, the taxonomic affinities of these fossils remain unclear (3). Eukaryotic fossils that can be assigned to extant taxonomic groups begin to appear around 1200 Ma (11) and become more widespread, abundant, and diverse in rocks ca. 800 Ma and younger (3, 12, 13).

Molecular estimation of divergence times has improved dramatically in recent years due the development of methods that incorporate uncertainty from sources that include phylogenetic reconstruction, fossil calibrations, and heterogeneous rates of molecular evolution (e.g., 1, 14, 15). Relaxed clock approaches account for heterogeneity in evolutionary rates across branches and enable the use of complex models of sequence evolution (reviewed in 16, 17), although debate continues as to the best method for relaxing the clock (18-20). The process of calibrating molecular clocks has also been greatly improved with the recognition that single

calibration points are insufficient (21-23), and current methods incorporate uncertainty from the fossil record by specifying calibration points as time distributions rather than points (16). Additional limitations in previous molecular clock studies of eukaryotes stem from the tradeoff between analyses of many taxa and calibration points but only a single gene (4) and analyses of many genes but a small number of taxa and calibrations (5, 24).

Molecular clock estimates rely on robust phylogenies. Reconstructions of relationships among major lineages of eukaryotes have begun to stabilize in recent years with the increasing availability of multigene data from diverse lineages (25-27). The majority of the >70 lineages of eukaryotes (28) fall within four major groups: Opisthokonta, Excavata, Amoebozoa, and SAR (Stramenopiles, Alveolates, and Rhizaria; 26, 27), while the placement of some photosynthetic lineages remains controversial (26, 29, 30). Greater data availability will also yield more accurate estimates of divergence times because more nodes are now available for calibration (31).

The availability of taxon- and gene-rich datasets coupled with flexible molecular clock methods make this an ideal time to revisit the timing of early eukaryotic evolution. Here, broadly sampled multigene trees are used to estimate dates, with rate heterogeneity across the tree and among genes incorporated into the model. We use 23 calibration points specified as prior distributions derived from fossils of Proterozoic and Phanerozoic age assigned to diverse lineages (Table 1). The Proterozoic fossil record is much more sparse (3, 8, 9), and the taxonomic assignment of some Proterozoic fossils has been called into question by a minority of researchers (e.g., 6). In the spirit of testing these ideas, we assess the impact of including of calibration constraints derived from Phanerozoic fossils alone and Phanerozoic plus Proterozoic fossils. We also assess divergence dates across analyses in which the phylogenetic tree varies by the position of the root, in the numbers of taxa included, and across different software platforms and models (Table 2).

Materials and Methods

Alignments

Alignments are derived from the 15 protein-coding genes analyzed in reference 23 (dataset '15:10'). Using this 88-taxon dataset as a starting point, taxa were added to capture additional lineages, particularly those with fossil data available (Table S1). Rapidly evolving taxa (e.g., *Encephalitozoon cuniculi*) and orphans (e.g., *Breviata anathema*) were removed to minimize rate heterogeneity for the clock analysis. The resulting 109-taxon data matrix includes 5696 characters, with each taxon having between 3 and 15 of the target genes (36% missing character data; Table S1; analyses a-c, e-p in Table 2). A 91-taxon alignment was created by removing additional taxa with either long branches or high levels of missing data to ensure that our results were not driven by these potential sources of artifact (analysis d).

Molecular dating analyses

Dating analyses were predominantly performed in BEAST v1.5.4 (32), and we also assessed results obtained in PhyloBayes 3.2f (33; see SI Text for analysis details). BEAST offers a number of desirable features, including flexible specification of prior distributions that enable the uncertainty of the fossil record to be realistically modeled, as well as the ability to co-estimate divergence times with topology, which may also produce better phylogenies (15). Although PhyloBayes allows exploration of a larger variety of molecular evolution and clock models, it requires calibration constraints to be specified as uniform distributions and a fixed topology (33). Only uncorrelated models are implemented in BEAST, and we ran our analyses with the uncorrelated lognormal model (UCL, see SI Text; 32). In PhyloBayes, we used the CIR autocorrelated model as it has been previously shown to provide a better fit for datasets with deep divergences (20) and the UGAM uncorrelated model as it is similar to UCL (33). There is much debate as to whether substitution rates are best modeled as autocorrelated across the tree or uncorrelated (15, 18-20). Autocorrelated models of the molecular clock

assume that evolutionary rates along a branch are dependent on the rate of the parent branch (16, 20), whereas uncorrelated models draw rates of evolution for each branch from a distribution of rates (15, 18). We compared divergence dates for eukaryotes obtained from different models to assess whether our conclusions were driven by the choice of a particular model (Fig. 1 and Table 2).

Calibration constraints

All calibration constraints (CC) incorporate error arising from age dating, stratigraphy and clade assignment when specifying the prior distribution (Table 1). Sixteen CCs were assigned based on fossils of Phanerozoic age and seven additional CCs were added from Proterozoic fossils (Table 1). The impact of these older fossils was assessed by analyzing the data with only the 16 Phanerozoic CCs ('Phan' analyses b,f,h,j,l,n,p) or with Phanerozoic and Proterozoic CCs ('All' analyses a,c-e,g,i,k,m,o). Calibration constraints were specified with prior distributions in BEAST using BEAUTi v1.5.4 (32) and were derived from a conservative reading of the fossil record (i.e., we err toward younger rather than older ages; SI text). Distributions were specified with long tails unless the fossil record provided minimum divergence information. Calibration constraints used for PhyloBayes were the same as in BEAST, but had to be specified as a uniform distribution (Table S2).

Assessing impact of the root on the inferred age of eukaryotes

Molecular clock analyses require a rooted tree. However the position of the eukaryotic root remains an open question; therefore, we compared age estimates from molecular clock analyses with multiple positions for the root of extant eukaryotes. First, the root was constrained the branch leading to the Opisthokonta or to Opisthokonta + Amoebozoa ('Unikonta') in accordance with current hypotheses (see SI text for discussion of the position of the eukaryotic root). In BEAST, the root was specified by constraining a monophyletic ingroup.

PhyloBayes requires the tree topology to be fixed, and we used the tree in Figure 2 rooted on either Opisthokonta or 'Unikonta'. Finally, for the third condition the root was estimated by the molecular clock criterion, as implemented in BEAST (SI Text), which yielded variable estimates of the location of the root.

Results

Taxon-rich analyses of multiple genes reveal a stability in divergence dates across the eukaryotic tree of life that is robust to changing taxon inclusion, the position of the root, molecular clock model, and choice of calibration points (Phanerozoic only or both Phanerozoic and Proterozoic fossils). Collectively, these analyses provide a mean age for the root of extant eukaryotes to between 1866 Ma and 1679 Ma in analyses including both Proterozoic and Phanerozoic calibrations ('All' analyses; Fig. 1*A* and Table 2). Varying the position of the root had little impact on estimated divergence dates across eukaryotes, especially for the estimated date of the root itself, which generally changed by less than 100 myr (Fig. 1*A*). Phylobayes estimates generally showed more uncertainty than those using BEAST, but around similar means. Similarly, estimates were robust to changing models (uncorrelated or autocorrelated) and to the inclusion of only Phanerozoic (Phan) or all calibrations (All) with one exception: under the autocorrelated CIR model estimates are much more recent in Phan analyses (1038 Ma and 1180 Ma; Fig. 1*A*).

Impact of calibration constraints on estimates of the origin of extant eukaryotes

We assessed the impact of including Proterozoic fossils, which are considered controversial by some (6, 7) by analyzing datasets without these seven calibration constraints (Phan analyses). In BEAST analyses, the exclusion of Proterozoic fossils shifted estimated divergence times toward the present, but not dramatically so: estimates for the mean age of root of extant eukaryotes fall between 1506-1471 Ma in Phan analyses (95% HPD range 1643-1347

Ma; Figs. 1*A*,S1,S5,S7 and Table 2, analyses b,f,h) as compared to 1837-1717 Ma (95% HPD range 1954-1601 Ma; Figs. 1*A*,2,S4,S6, analyses a,e,g,) when Proterozoic fossils were included (All analyses). Similar dates were recovered in Phan and All PhyloBayes analyses when the UGAM model (uncorrelated) of the molecular clock was assumed (Fig. 1*A* and Table 2, analyses i-l).

It is important to note that of the seven Proterozoic calibration points used in our analyses, only the *Bangiomorpha* point is controversial in terms of either systematic attribution or age. The *Bangiomorpha* calibration constraint is more than 400 million years (myr) older than our other Proterozoic constraints (Table 1). To determine whether this calibration point drives our results in analyses with All calibrations, we assessed the age of the root with a much more conservative estimate for the age of this red alga at 720 Ma ('All 720'; Table 2, analysis c). A number of factors place the age of Bangiomorpha around 1200 Ma (see SI Text); however, given the importance of the fossil we also assigned an age of 720 Ma to this constraint, representing the absolute younger bound of the Hunting Formation, Canada, in which it is found (SI Text; 11). In BEAST, placing the *Bangiomorpha* constraint at 720 Ma shifted the estimated age of the root by only 95 myr toward the present (Figs. 1*A* and S3, analysis c).

The autocorrelated CIR model combined with the low number of substitutions on deep branches of the eukaryotic tree appears to be more sensitive to the distribution of calibration dates included in these analyses. Under the CIR autocorrelated model a consistent age was also estimated with All calibrations included (1798 – 1691 Ma; Fig. 1*A*, analyses m,o), although confidence intervals are in general greater in PhyloBayes analyses (Fig. 1*A*, analyses i-p). However, excluding Proterozoic calibration points did cause estimated ages to shift more than 600 myr younger under the CIR model (1180 – 1038 Ma; Fig. 1*A*, analyses n,p), pushing the estimated age for the root of extant eukaryotes younger than the widely accepted date for the *Bangiomorpha* fossils. Similarly, the CIR analyses in PhyloBayes were sensitive to the age of the *Bangiomorpha* constraint, and shifted more than 500 myr younger to 1296 Ma and 1167 Ma

in analyses with All calibration points and the *Bangiomorpha* constraint set to 720 Ma rooted with Opisthokonta and 'Unikonta' respectively (Dataset S1). The necessity of using PhyloBayes to explore the differences between autocorrelated and uncorrelated models introduces confounding factors, as PhyloBayes requires both uniform distributions around calibration points and a fixed tree topology. Given that calibration points are likely best represented by more informative distributions and that the topology of the tree is not fully known, we focus the rest of our discussions on the results from BEAST, although data from all PhyloBayes analyses are available in Figure 1*A* and Dataset S1.

Origin of major clades

In most analyses, the major clades of extant eukaryotes diverged prior to 1200 Ma, with the major clades SAR, Excavata and Amoebozoa arising within a similar time frame, as evidenced by overlapping 95% highest probability density ranges (HPD, akin to confidence intervals; Figs. 1, 2 and S1-7 and Dataset S1). The 95% HPD intervals are wider for clades with few calibration points, such as Excavata and Amoebozoa (Fig. 1*B*). Estimates for the last common ancestor of extant Opisthokonta are younger than the other clades, at 1389 –1240 Ma in analyses with 'All' calibration constraints.

Exclusion of Proterozoic calibration constraints shifted age estimates for the origins of major extant eukaryotic clades younger by 200 to 300 myr (Fig. 1*B*). Differences in divergence times are relatively small for nested clades, e.g., the 95% HPD for Alveolata shifts from 1445-1236 Ma in analysis a (Fig. 2) to 1206-1020 Ma with only Phanerozoic calibration points (analysis b; Fig. S1). Not surprisingly, the differing calibration schemes had their most dramatic impact on the estimated age of the red algae, which changes from 1285-1180 Ma 95% HPD (Fig. 2) to 959-625 Ma 95% HPD when Proterozoic calibration points, including the constraint on red algae at 1174 Ma in accordance with the widely cited age for *Bangiomorpha*, are excluded

(Fig. S1). Estimated ages of major clades were also much younger in analyses using the CIR model with Phan calibrations (analyses n,p; Dataset S1).

The topology of the eukaryotic tree produced through co-estimation of phylogeny and divergence times in BEAST is broadly consistent with other analyses (SI Text; 26, 27). Hence, the BEAST topology was also used for the PhyloBayes analyses, which require a fixed topology. While the relationships among the photosynthetic eukaryotes remain uncertain (e.g., 26), our analyses suggest that many photosynthetic clades, including red and green algae, diverged within a similar time frame (Fig. 2). These results imply an early acquisition of photosynthesis in eukaryotes, in accordance with previous molecular clock estimates (34) and the ca. 1200 Ma age assigned to the red algal fossil *Bangiomorpha* (11).

Discussion

When both Phanerozoic and Proterozoic fossils are considered, the molecular clock analyses presented here suggest that the last common ancestor of extant eukaryotes lived between 1866 and 1679 Ma. We favor these more inclusive analyses as they should reveal a more accurate picture of eukaryotic diversification, especially since the chosen fossils are widely accepted by paleontologists and prior distributions were assigned in a conservative manner that accounts for age uncertainties. Estimated ages are younger when we remove Proterozoic calibration constraints, but not dramatically so with the notable exception of the autocorrelated model CIR as implemented in PhyloBayes with only Phanerozoic calibrations. Thus, our results tend to place the last common ancestor of extant eukaryotes deep within the Proterozoic Eon.

Our estimates for the timing of the origin of extant eukaryotes are in line with fossil evidence (3, 13), but reject the hypothesis that eukaryotes originated only 850 million years ago (6, 7). Fossils provide minimum dates, leaving open the possibility that clades evolved much earlier than the first fossil appearance (e.g. 3, 35). Thus, it is not surprising that divergence times for many eukaryotic clades are older than their first unambiguous fossil occurrence (Table

3). The paleontological literature contains some references to eukaryotic fossils older than our estimate of the last common ancestor. In some cases, these paleontological reports are incorrect or ambiguous. For example, large carbonaceous fossils assigned to the genus *Grypania* were originally reported to be older than our molecular clock estimate (36), but more recent radiometric dates indicate an age of 1874 ± 9 Ma (37), consistent with the clock analyses presented here. Older still are the 50-300 μm spheroidal microfossils described from ca. 3200 Ma rocks by Javaux et al. (38; proposed as possible eukaryotes by, 39) and sterane biomarkers from 2700 Ma shales (2). Whether or not these materials record Archean eukaryotes remains a subject of debate (38, 40). Our molecular clock estimates suggest that if these fossils do represent eukaryotes, they record stem lineages—early and now representatives of eukaryotic groups that diverged prior to the last common ancestor of extant members.

The major lineages of extant eukaryotes (Opisthokonta, SAR, Excavata and Amoebozoa) are projected to have diverged from one another by the Mesoproterozoic Era (1600 to 1000 Ma), relatively early in the history of the domain (Fig. 1 and Table 3). This, in turn, suggests that these lineages were present for hundreds of millions of years before the observed increase in the abundance and diversity of eukaryotic microfossils beginning roughly 800 Ma (3, 41-44). Our molecular clock estimates indicate that stem groups were present well before recognizable members of crown lineages—monophyletic groups consisting of the last common ancestor of living representatives and its descendants—diversified. A similar pattern of long stems preceding diversification is seen in animal and plants and may be a consistent pattern in evolution (42).

Fossils and our molecular clock analyses agree that eukaryotes originated and diversified during a time when oceans differed substantially from the modern seas. Increasingly, geochemical data indicate that for much of the Proterozoic Eon, mildly oxic surface waters lay above an oxygen minimum zone that was persistently anoxic and commonly sulfidic (45, 46). Such conditions are compatible with scenarios for eukaryogenesis that rely on

anaerobic methanogens in symbiotic partnership with facultatively aerobic proteobacteria or sulfate reducers (see refs in 47), as facultatively anaerobic mitochondria may have enabled early eukaryotes to live in the sulfidic Proterozoic oceans (48). As sulfide interferes with the function of mitochondria to aerobically respiring eukaryotes, the radiation of diverse species within eukaryotic clades may have become possible only as sulfidic subsurface waters began to wane about 800 Ma (49). Alternatively, early eukaryotic evolution may have occurred in coastal environments sheltered from the impact of sulfidic waters or in freshwater systems, which are both poorly sampled by the geologic record and not impacted by sulfidic oceanic water masses (50). Consistent with this view, moderately diverse assemblages of fossil eukaryotes occur well ventilated lake deposits of the 1200-900 Ma Torridonian succession, Scotland (51,52), and in coastal marine deposits of the ca.1400-1500 Ma Roper Group, Australia (53).

Within Proterozoic oceans, low concentrations of biologically available nitrogen may also have inhibited the diversification of photosynthetic eukaryotes (54). Many cyanobacteria and other photosynthetic bacteria are capable of nitrogen fixation, ameliorating the impact of nitrate and ammonia limitation on primary production. Eukaryotes, however, have no such capacity; thus, it may not be a coincidence that biomarkers indicating an expanding importance of algae in marine primary production occur in conjunction with geochemical data recording the spread of oxygen through later Neoproterozoic oceans (55). In our analyses, the clade that contains extant photosynthetic taxa, including green algae plus land plant and red algae, arose between 1670 and 1428 Ma (Table 3), but diversification within these lineages occurred later in the Neoproterozoic and may correspond to a changing redox profile in the oceans (e.g. Fig. 2).

Discrepancy between these and previous molecular clock studies

Previous molecular clock studies yielded vastly different dates for the root of extant eukaryotes, ranging from 1100 Ma to 3970 Ma (1). In a recent analysis of SSU-rDNA from 83 broadly sampled eukaryotes, Berney and Pawlowski (4) placed the origin of eukaryotes at 1100

Ma, a conclusion that was robust to changing the position of the root (Table S2 in Ref. 4). They had numerous Phanerozoic calibration constraints specified as either minimum or maximum divergence dates (4), but they found that including Proterozoic calibration points, such as *Bangiomorpha* at 1200 Ma, shifted their estimates of the origin and diversification of eukaryotes by 1000 to 2500 Ma (Table 1 in Ref. 4). The age discrepancy observed by Berney and Pawlowski (4) when Proterozoic calibration constraints are included contrasts sharply with the relative stability of dates seen in our analyses (Table 2). We hypothesize that the increased gene and taxon sampling as well as the use of flexible prior distributions of calibration points as implemented in BEAST are major factors contributing to the stability of molecular clock estimation in our analyses.

Conclusion

Our molecular clock analyses yield a timeline of eukaryotic evolution that is congruent with the paleontological record and robust to varying analytical conditions. According to our analyses, crown (extant) groups of eukaryotes arose in the Paleoproterozoic Era (2500-1600 Ma) and began to diversify soon thereafter, suggesting that early eukaryotic evolution was influenced by anoxic and sulfidic water masses in contemporaneous oceans. The stability in our analysis across a range of variables is a welcome departure from the large age discrepancies reported in earlier molecular analyses, reflecting improved paleontological interpretation, advancements in molecular methods, and the rapidly growing body of molecular data from diverse eukaryotes.

Acknowledgements

Thanks to Ben Normark, Rob Dorit and Sam Bowser for useful discussions. Thanks to Jeff Thorne (North Carolina State University, USA) and Bengt Sennblad (Karolinska Institutet, Stockholm Bioinformatics Center and SciLifeLab, Stockholm, Sweden) for helpful discussions about molecular clock models. This manuscript has been improved following the comments of Emmanuelle Javaux, Andrew Roger, and Heroen Verbruggen. We thank Jessica Grant for help in developing the dataset. Many thanks also to Tony Caldanaro at Smith College for technical help in running the analyses. This research was supported by a grant from the NASA Astrobiology Institute to AHK, and by NSF Assembling the Tree of Life (043115) and NSF RUI Systematics (0919152) awards to LAK. DJGL is supported by CNPq-Brazil, GDE Fellowship #200853/2007-4.

References

1. Roger AJ & Hug LA (2006) The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philos T Roy Soc B* 361:1039-1054.
2. Brocks JJ, Logan GA, Buick R, & Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes. *Science* 285(5430):1033-1036.
3. Knoll AH, Javaux EJ, Hewitt D, & Cohen P (2006) Eukaryotic organisms in Proterozoic oceans. *Philos Trans R Soc B* 361(1470):1023-1038.
4. Berney C & Pawlowski J (2006) A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proc Roy Soc B* 273(1596):1867-1872.
5. Douzery EJP, Snell EA, Bapteste E, Delsuc F, & Philippe H (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci, USA* 101(43):15386-15391.
6. Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *Int J Syst Evol Microbiol* 52:297-354.
7. Cavalier Smith T (2010) Deep phylogeny, ancestral groups and the four ages of life. *Philos Trans R Soc B* 365(1537):111-132.
8. Porter SM (2004) The fossil record of early eukaryotic diversification. *Paleontological Society Papers* 10:35-50.
9. Javaux EJ, Knoll AH, & Walter M (2003) Recognizing and interpreting the fossils of early eukaryotes. *Origins Life Evol B* 33(1):75-94.
10. Javaux EJ, Knoll AH, & Walter MR (2004) TEM evidence for eukaryotic diversity in mid-Proterozoic oceans. *Geobiology* 2:121-132.
11. Butterfield NJ (2000) *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiol* 26(3):386-404.
12. Porter SM, Meisterfeld R, & Knoll AH (2003) Vase-shaped microfossils from the Neoproterozoic Chuar Group, Grand Canyon: A classification guided by modern testate amoebae. *J Paleontol* 77(3):409-429.
13. Javaux EJ (2007) The early eukaryotic fossil record. *Eukaryotic Membranes and Cytoskeleton: Origins and Evolution,* Advances in Experimental Medicine and Biology), Vol 607, pp 1-19.
14. Welch JJ & Bromham L (2005) Molecular dating when rates vary. *Trends Ecol Evol* 20(6):320-327.
15. Drummond AJ, Ho SYW, Phillips MJ, & Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4(5):699-710.
16. Ho SYW & Phillips MJ (2009) Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst Biol* 58(3):367-380.
17. Rutschmann F (2006) Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Divers Distrib* 12:35-48.
18. Linder M, Britton T, & Sennblad B (2011) Evaluation of Bayesian models of substitution rate evolution—parental guidance versus mutual independence. *Syst. Biol.* 60(3):329-342.
19. Ho SYW (2009) An examination of phylogenetic models of substitution rate variation among lineages. *Biology Letters* 5:421-424.

20. Lepage T, Bryant D, Philippe H, & Lartillot N (2007) A general comparison of relaxed molecular clock models. *Mol Biol Evol* 24(12):2669-2680.
21. Graur D & Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet* 20(2):80-86.
22. Hug LA & Roger AJ (2007) The impact of fossils and taxon sampling on ancient molecular dating analyses. *Mol Biol Evol* 24(8):1889-1897.
23. Forest F (2009) Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. *Ann Bot* 104(5):789-794.
24. Hedges SB, Blair JE, Venturi ML, & Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* 4.
25. Adl SM*, et al.* (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Euk Microbiol* 52(5):399-451.
26. Parfrey LW*, et al.* (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst Biol* 59(5):518-533.
27. Hampl V*, et al.* (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci, USA* 106(10):3859-3864.
28. Patterson DJ (1999) The diversity of eukaryotes. *Am Nat* 154:S96-S124.
29. Lane CE & Archibald JM (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol Evol* 23(5):268-275.
30. Baurain D*, et al.* (2010) Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol* 27(7):1698-1709.
31. Heath TA, Hedtke SM, & Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. *J Syst Evol* 46(3):239-257.
32. Drummond A & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7(1):214.
33. Lartillot N, Lepage T, & Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17):2286-2288.
34. Yoon HS, Hackett JD, Ciniglia C, Pinto G, & Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. (Translated from English) *Mol Biol Evol* 21(5):809-818 (in English).
35. Donoghue PCJ & Benton MJ (2007) Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol Evol* 22(8):424-431.
36. Han TM & Runnegar B (1992) Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunee iron-formation, Michigan. *Science* 257(5067):232-235.
37. Schneider DA, Bickford ME, Cannon WF, Schulz KJ, & Hamilton MA (2002) Age of volcanic rocks and syndepositional iron formations, Marquette Range Supergroup: Implications for the tectonic setting of Paleoproterozoic iron formations of the Lake Superior. *Can J Earth Sci* 39:999-1012.
38. Javaux EJ, Marshall CP, & Bekker A (2010) Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature* 463(7283):934-939.
39. Buick R (2010) Ancient life: early acritarchs. *Nature* 463:885-886.
40. Rasmussen B, Fletcher IR, Brocks JJ, & Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* 455(7216):1101-U1109.
41. Knoll AH (1994) Proterozoic and early Cambrian protists: evidence for accelerating evolutionary tempo. *Proc Natl Acad Sci, USA* 91:6743-6750.

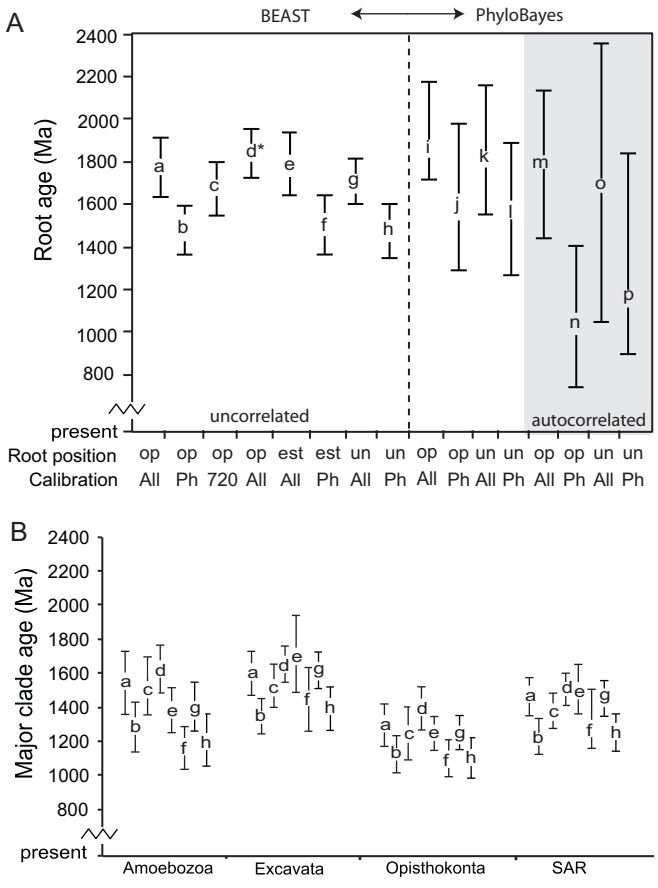
42. Knoll AH (2011) The multiple origins of complex multicellularity. *Annu Rev Earth Pl Sci* 39:217-239.
43. Yin L & Yuan X (2007) Radiation of Meso-Neoproterozoic and early Cambrian protists inferred from the microfossil record of China. *Palaeogeogr Palaeocl* 254:350-361.
44. Porter SM (2006) Heterotrophic Eukaryotes. *Neoproterozoic Geobiology and Paleobiology*, eds Xiao S & Kaufman AJ (Springer, Netherlands), pp 1-21.
45. Canfield DE (1998) A new model for Proterozoic ocean chemistry. *Nature* 396(6710):450-453.
46. Johnston DT, Wolfe-Simon F, Person A, & Knoll AH (2009) Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proc Natl Acad Sci, USA* 106:16925-16929.
47. Embley TM & Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440(7084):623-630.
48. Mentel M & Martin W (2008) Energy metabolism among eukaryotic anaerobes in light of Proterozoic ocean chemistry. *Philos Trans R Soc B* 363(1504):2717-2729.
49. Johnston DT*, et al.* (2010) An emerging picture of Neoproterozoic ocean chemistry: Insights from the Chuar Group, Grand Canyon, USA. *Earth Planet Sc Lett* 290:64-73.
50. Cavalier Smith T (2009) Megaphylogeny, cell body plans, adaptive zones: causes and timing of eukaryote basal radiations *J Euk Microbiol* 56(1):26-33.
51. Parnell J, Boyce AJ, Mark D, Bowden S, & Spinks S (2010) Early oxygenation of the terrestrial environment during the Mesoproterozoic. *Nature* 468(7321):290-293.
52. Strother P, Battison L, Brasier MD, & Wellman C (2011) Earth's earliest non-marine eukaryotes. *Nature*.
53. Javaux EJ, Knoll AH, & Walter MR (2001) Morphological and ecological complexity in early eukaryotic ecosystems. *Nature* 412(6842):66-69.
54. Anbar AD & Knoll AH (2002) Proterozoic ocean chemistry and evolution: A bioinorganic bridge? *Science* 297(5584):1137-1142.
55. Knoll AH, Summons RE, Waldbauer JR, & Zumberge J (2007) The geological succession of primary producers in the oceans. *The Evolution of Primary Producers in the Sea*, eds Falkowski PG & Knoll AH (Elsevier, Burlington), pp 133-163.
56. Smithson TR & Rolfe WDI (1990) *Westlothiana* Gen. nov. - naming the earliest known reptile *Scottish J Geol* 26:137-138.
57. Crane PR, Friis EM, & Pedersen KR (1995) The origin and early diversification of angiosperms *Nature* 374(6517):27-33.
58. Taylor TN, Hass T, & Kerp H (1999) The oldest fossil ascomycetes. *Nature* 399(6737):648-648.
59. Bown PR (1998) *Calcareous nannofossil biostratigraphy* (Kluwer Academic Publishers, London) p 328.
60. Harwood DM, Nikolaev VA, & Winter DM (2007) Cretaceous records of diatom evolution, radiation, and expansion. *Pond Scum to Carbon Sink: Geological and Environmental Applications of the Diatoms, Paleontological Society Short Course, October 27, 2007*, ed Starratt S (Paleontological Society Papers), Vol 13, pp 33-59.
61. Fensome RA, Saldarriaga JF, & Taylor F (1999) Dinoflagellate phylogeny revisited: reconciling morphological and molecular based phylogenies. *Grana* 38(2-3):66-80.

62. Rubinstein CV, Gerrienne P, de la Puente GS, Astini RA, & Steemans P (2010) Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytologist* 188(2):365-369.
63. Dostál O & Prokop J (2009) New fossil insects (Diaphanopterodea: Martynoviidae) from the Lower Permian of the Boskovice Basin, southern Moravia. *GeoBios* 42(4):495-502.
64. Friis EM, Pedersen KR, & Crane PR (2010) Diversity in obscurity: fossil flowers and the early history of angiosperms. *Philos Trans R Soc B* 365(1539):369-382.
65. Sun G, Dilcher D, Wang H, & Chen Z (2011) A eudicot from the Early Cretaceous of China. *Nature* 471:625-628.
66. Gray J & Boucot AJ (1989) Is *Moyeria* a euglenoid? *Lethaia* 22(4):447-456.
67. McIlroy D, Green OR, & Brasier MD (2001) Palaeobiology and evolution of the earliest agglutinated Foraminifera: Platysolenites, Spirosolenites and related forms. *Lethaia* 34(1):13-29.
68. Hua H, Chen Z, Yuan XL, Xiao SH, & Cai YP (2010) The earliest Foraminifera from southern Shaanxi, China. *Sci China-Earth Sci* 53(12):1756-1764.
69. Kooistra W, Gersonde R, Medlin L, & Mann DG (2007) The Origin and Evolution of the Diatoms: Their Adaptation to a Planktonic Existence. *The Evolution of Primary Producers in the Sea*, eds Falkowski PG & Knoll AH (Elsevier, Burlington), pp 201-249.
70. Lipps HJ (1993) *Fossil Prokaryotes and Protists* (Blackwell Scientific Publications, Boston).
71. Kenrick P & Crane PR (1997) The origin and early evolution of plants on land. *Nature* 389(6646):33-39.
72. Shu DG*, et al.* (1999) Lower Cambrian vertebrates from South China. *Nature* 402(6757):42-46.
73. Love GD*, et al.* (2009) Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* 457(7230):718-U715.
74. Cohen PA, Knoll AH, & Kodner RB (2009) Large spinose microfossils in Ediacaran rocks as resting stages of early animals. *Proc Natl Acad Sci, USA* 106(16):6519-6524.
75. Martin MW*, et al.* (2000) Age of Neoproterozoic bilatarian body and trace fossils, White Sea, Russia: Implications for metazoan evolution. *Science* 288(5467):841-845.
76. Butterfield NJ, Knoll AH, & Swett K (1994) Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. *Fossils and Strata* 34:1-84.
77. Summons RE & Walter MR (1990) Molecular fossils and microfossils of prokaryotes and protists from Proterozoic sediments. *Am J Sci* 290-A:212-244.
78. Xiao SH, Knoll AH, Yuan XL, & Pueschel CM (2004) Phosphatized multicellular algae in the Neoproterozoic Doushantuo Formation, China, and the early evolution of florideophyte red algae. *Am J Bot* 91(2):214-227.

Figure legends

Figure 1. Summary of mean divergence dates for the most recent common ancestor of major clades of extant eukaryotes. Letters are at the mean divergence time and denote analyses, as detailed in Table 2. Error bars represent 95% highest posterior density (HPD) for BEAST analyses (a-h) and the 95% confidence interval for PhyloBayes (i-p). (*A*) Estimated age of the root of extant eukaryotes across analyses. An uncorrelated molecular clock model was used for all analyses except those in the grey box. Root position: Opis = root constrained to Opisthokonta; Uni = root constrained to 'Unikonta'; Estim = root estimated by BEAST. Calibration: All = all Phanerozoic and Proterozoic CCs; Phan = Phanerozoic CCs only; 720 = All CCs with the minimum age of red algae set to 720 Ma. d* = 91 taxa. (*B*) Estimated ages of major clades from BEAST analyses.

Figure 2. Time calibrated tree of extant eukaryotes using All calibration points, 109 taxa, and root constrained to Opisthokonta. Nodes are at mean divergence times and grey bars represent 95% HPD of node age. Geological time scale is on top and absolute time scale is shown on bottom in Ma. Thick vertical bars demarcate Eras and thin vertical lines denote Periods, with dates derived from the 2009 International Stratigraphic Chart. ● = Node calibrated with Phanerozoic fossils, ○= Node calibrated with Proterozoic fossils. Note that estimated ages of calibrated nodes differ from the prior calibration constraints (Table 1) because they have been modified by sequence data.


A
BEAST
PhyloBayes
Root age (Ma)
2400
2200
2000
1800
1600
1400
1200
1000
800
present
uncorrelated
autocorrelated
Root position
Calibration
op All
op Ph
op 720
op All
est All
est Ph
un All
un Ph
op All
op Ph
un All
un Ph
op All
op Ph
un All
un Ph
B
Major clade age (Ma)
Amoebozoa
Excavata
Opisthokonta
SAR

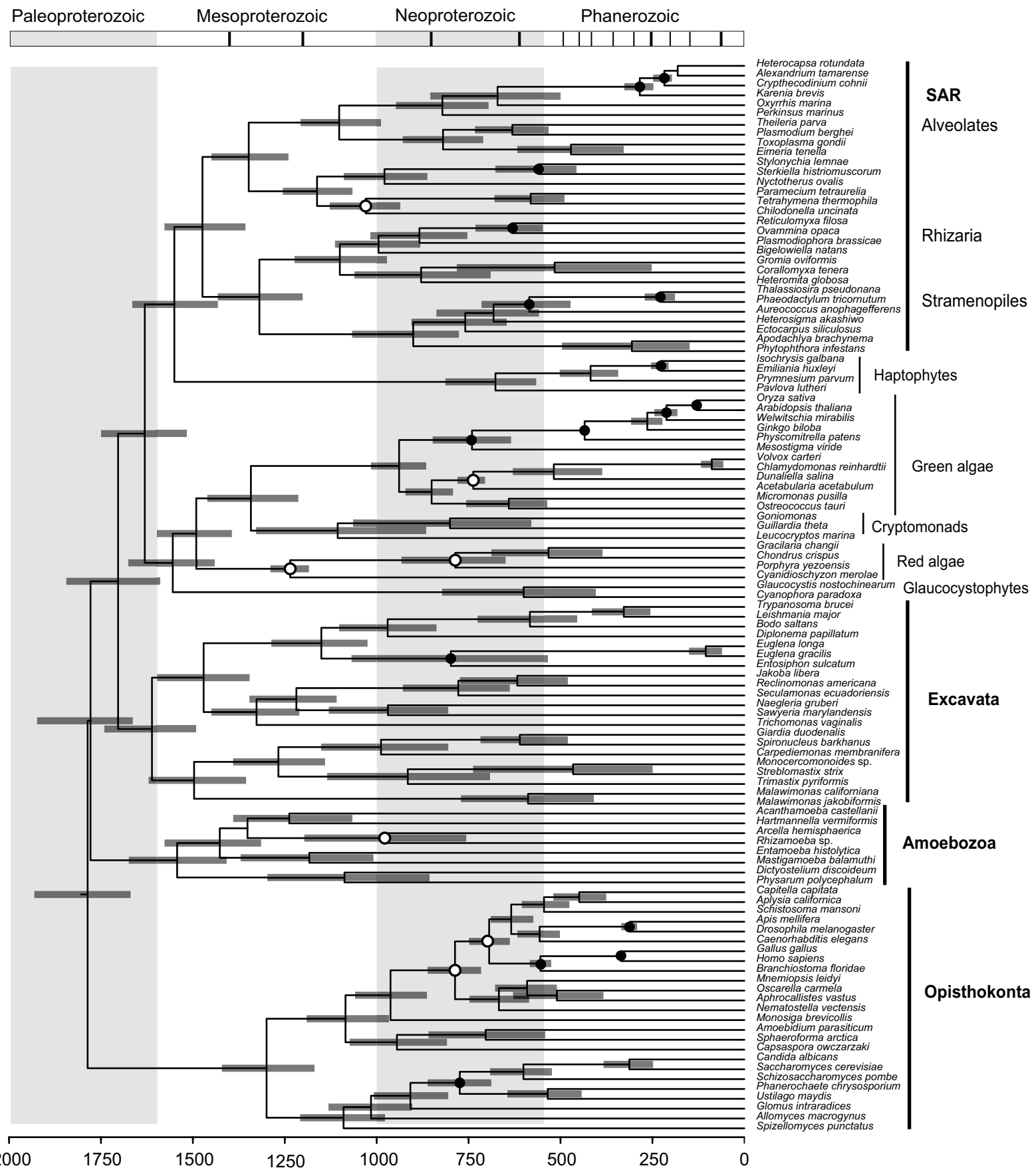

Paleoproterozoic
Mesoproterozoic
Neoproterozoic
Phanerozoic
Heterocapsa rotundata
Alexandrium tamarense
Crypthecodinium cohnii
Karenia brevis
Oxyrrhis marina
Perkinsus marinus
Theileria parva
Plasmodium berghei
Toxoplasma gondii
Eimeria tenella
Stylonychia lemnae
Sterkiella histriomuscorum
Nyctotherus ovalis
Paramecium tetraurelia
Tetrahymena thermophila
Chilodonella uncinata
Reticulomyxa filosa
Ovammina opaca
Plasmodiophora brassicae
Bigelowiella natans
Gromia oviformis
Corallomyxa tenera
Heteromita globosa
Thalassiosira pseudonana
Phaeodactylum tricornutum
Aureococcus anophagefferens
Heterosigma akashiwo
Ectocarpus siliculosus
Apodachlya brachynema
Phytophthora infestans
Isochrysis galbana
Emiliania huxleyi
Prymnesium parvum
Pavlova lutheri
Oryza sativa
Arabidopsis thaliana
Welwitschia mirabilis
Ginkgo biloba
Physcomitrella patens
Mesostigma viride
Volvox carteri
Chlamydomonas reinhardtii
Dunaliella salina
Acetabularia acetabulum
Micromonas pusilla
Ostreococcus tauri
Goniomonas
Guillardia theta
Leucocryptos marina
Gracilaria changii
Chondrus crispus
Porphyra yezoensis
Cyanidioschyzon merolae
Glaucocystis nostochinearum
Cyanophora paradoxa
Trypanosoma brucei
Leishmania major
Bodo saltans
Diplonema papillatum
Euglena longa
Euglena gracilis
Entosiphon sulcatum
Jakoba libera
Reclinomonas americana
Seculamonas ecuadoriensis
Naegleria gruberi
Sawyeria marylandensis
Trichomonas vaginalis
Giardia duodenalis
Spironucleus barkhanus
Carpediemonas membranifera
Monocercomonoides sp.
Streblomastix strix
Trimastix pyriformis
Malawimonas californiana
Malawimonas jakobiformis
Acanthamoeba castellanii
Hartmannella vermiformis
Arcella hemisphaerica
Rhizamoeba sp.
Entamoeba histolytica
Mastigamoeba balamuthi
Dictyostelium discoideum
Physarum polycephalum
Capitella capitata
Aplysia californica
Schistosoma mansoni
Apis mellifera
Drosophila melanogaster
Caenorhabditis elegans
Gallus gallus
Homo sapiens
Branchiostoma floridae
Mnemiopsis leidyi
Oscarella carmela
Aphrocallistes vastus
Nematostella vectensis
Monosiga brevicollis
Amoebidium parasiticum
Sphaeroforma arctica
Capsaspora owczarzaki
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Phanerochaete chrysosporium
Ustilago maydis
Glomus intraradices
Allomyces macrogynus
Spizellomyces punctatus
SAR
Alveolates
Rhizaria
Stramenopiles
Haptophytes
Green algae
Cryptomonads
Red algae
Glaucocystophytes
Excavata
Amoebozoa
Opisthokonta
2000
1750
1500
1250
1000
750
500
250
0

Table 1. Calibration constraints for dating the eukaryotic tree of life

Taxon	Fossil	Eon[1]	Calibration[2] min	Calibration[2] dist	Refs
Amniota	*Westlonthania*	Phan	328.3	4,3	(56)
Angiosperms	Oldest angio pollen	Phan	133.9	2,10	(57)
Ascomycetes	*Paleopyrenomycites*	Phan	400	4,50	(58)
Coccolithophores	Earliest Heterococcolith	Phan	203.6	2,8	(59)
Diatoms	Earliest diatoms	Phan	133.9	2,100	(60)
Dinoflagellates	Earliest gonyaulacales	Phan	240	2,10	(61)
Embryophytes	Land plant spores	Phan	471	2,20	(62)
Endopterygota	Mecoptera	Phan	284.4	5,5	(63)
Eudicots	Eudicot pollen	Phan	125	2,1.5	(64, 65)
Euglenids	*Moyeria*	Phan	450	2,40	(66)
Foraminifera	Oldest forams	Phan	542	2,200	(67, 68)
Gonyaulacales	Gonyaulacaceae split	Phan	196	2,10	(61)
Pennate diatoms	Oldest pennate	Phan	80	3,5	(69)
Spirotrichs	Oldest tintinnids	Phan	444	2.5,100	(70)
Trachaeophytes	Earliest trachaeophytes	Phan	425	4,2.5	(71)
Vertebrates	*Haikouichthys*	Phan	520	3,5	(72)
Animals	LOEMs, sponge biomarkers	Protero	632	2,300	(73, 74)
Arcellinida	*Paleoarcella*	Protero	736	2,300	(12)
Bilateria	*Kimberella*	Protero	555	2,30	(75)
Chlorophytes	*Palaeastrum*	Protero	700	2.5,300	(76)
Ciliates	Gammacerane	Protero	736	2.5,300	(77)
Florideophyceae	Doushantuo red algae	Protero	550	2.5,100	(78)
Red algae[3]	*Bangiomorpha*	Protero	1174	3,250	(11)

[1]Eon: Phan = Phanerozoic, Protero. = Proterozoic, Proterozoic calibrations are excluded from Phan analyses.
[2]Calibration constraints are specified for BEAST using a gamma distribution with a minimum date in Ma based on the fossil record parameters as indicated: min = minimum divergence data; dist = gamma prior distribution (shape, scale). See Table S2 for details of PhyloBayes calibrations.
[3]In the All720 analysis (c) the minimum age constraint for the red algae node is 720 Ma.

Table 3. Comparison of major node ages when all calibration constraints are used to fossil dates

Major clade	Estimated age	Oldest fossil	Ref
Eukaryotes	*	1800	(3)
Extant eukaryotes	1679 - 1866	1200	(11)
Amoebozoa	1384 - 1624	800	(12)
Excavata	1510 - 1699	450	(66)
Opisthokonta	1240 - 1481	632	(74)
Photosynthetic clade	1428 - 1670	1200	(11)
Rhizaria	1017 - 1256	550	(67, 68)
SAR	1365 - 1577	736	(77)

Ages are in Ma. Estimated age is range of mean dates from 'All' analyses. *The age of the root of all eukaryotes is not estimated because molecular clock studies can only inform the timing of extant clades.

Table 2. Estimates of dates for the last common ancestor of extant eukaryotes across analyses

Analysis	Taxa	CCs	Root	Root age (Ma) mean	range	model	Program	Tree
a	109	All	Opis	1774	1632 - 1911	UCL	BEAST	Fig. 2
b	109	Phan	Opis	1478	1362 - 1595	UCL	BEAST	Fig. S1
c	109	All 720	Opis	1679	1548 - 1797	UCL	BEAST	Fig. S2
d	91	All	Opis	1837	1725 - 1954	UCL	BEAST	Fig. S3
e	109	All	Estim	1784	1639 - 1939	UCL	BEAST	Fig. S4
f	109	Phan	Estim	1506	1365 - 1643	UCL	BEAST	Fig. S5
g	109	All	Uni	1717	1601 - 1819	UCL	BEAST	Fig. S6
h	109	Phan	Uni	1471	1347 - 1604	UCL	BEAST	Fig. S7
i	109	All	Opis	1866	1569 - 2235	UGAM	PhyloBayes	-
j	109	Phan	Opis	1594	1288 - 1979	UGAM	PhyloBayes	-
k	109	All	Uni	1810	1549 - 2161	UGAM	PhyloBayes	-
l	109	Phan	Uni	1561	1268 - 1886	UGAM	PhyloBayes	-
m	109	All	Opis	1798	1441 - 2133	CIR	PhyloBayes	-
n	109	Phan	Opis	1038	889 - 1350	CIR	PhyloBayes	-
o	109	All	Uni	1691	1048 - 2357	CIR	PhyloBayes	-
p	109	Phan	Uni	1180	897 - 1839	CIR	PhyloBayes	-

CCs = Calibration constraints. Phan = calibration points of Phanerozoic age included. All = 22 calibration points of Phanerozoic and Proterozoic age included. All 720 = *Bangiomorpha* CC set to 720 Ma. Root = position of the root: Opis = root constrained to Opisthokonta; Uni = root constrained to 'Unikonta'; Estim = root estimated by BEAST. Model = molecular clock model: UCL = uncorrelated log normal; UGAM = uncorrelated gamma model; CIR = autocorrelated CIR model. Root age range is the 95% HPD for BEAST analyses and min and max ages of 95% confidence interval for PhyloBayes. See Table S1 for details of taxon sampling and Table 1 for calibration constraints. All trees are available in the Dataset S1.