A site-and-branch-heterogeneous model on an expanded dataset favor mitochondria as sister to known Alphaproteobacteria

Sergio A Muñoz-Gómez, Edward Susko, Kelsey Williamson, Laura Eme, Claudio Slamovits, David Moreira, Purificación López-García

A site-and-branch-heterogeneous model on an expanded dataset favor mitochondria as sister to known Alphaproteobacteria

Sergio Munoz-Gomez (smunozgo@asu.edu)
Arizona State University
Edward Susko
Dalhousie University
Kelsey Williamson
Dalhousie University
Laura Eme
Université Paris-Saclay
Claudio Slamovits
Dalhousie University
David Moreira
Université Paris-Saclay
Purificacion Lopez-Garcia
Université Paris-Saclay
Andrew Roger
Dalhousie University https://orcid.org/0000-0003-1370-9820

Article

Keywords: Alphaproteobacteria, mitochondria, evolutionary genetics

Posted Date: May 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-557223/v1

Title: A site-and-branch-heterogeneous model on an expanded dataset favors mitochondria as sister to known *Alphaproteobacteria*

Authors: Sergio A. Muñoz-Gómez[1,4]*, Edward Susko[2], Kelsey Williamson[1], Laura Eme[3], Claudio H. Slamovits[1], David Moreira[3], Purificación López-García[3], and Andrew J. Roger[1]*

Affiliations:

[1]Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada.
[2]Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada.
[3]Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France.
[4]Current affiliation: Ecologie Systématique Evolution, Université Paris-Saclay, AgroParisTech, Orsay, France.
*Correspondence to: sergio.munoz@universite-paris-saclay.fr and andrew.roger@dal.ca

Abstract

Determining the phylogenetic origin of mitochondria is key to understanding the ancestral mitochondrial symbiosis and its role in eukaryogenesis. However, the precise evolutionary relationship between mitochondria and their closest bacterial relatives remains hotly debated. The reasons include pervasive phylogenetic artefacts, as well as limited protein and taxon sampling. Here, we developed a new model of protein evolution that accommodates both across-site and across-branch compositional heterogeneity. We applied this site-and-branch-heterogeneous model (MAM60+GFmix) to a considerably expanded dataset that comprises 108 mitochondrial proteins of alphaproteobacterial origin, and novel metagenome-assembled genomes from microbial mats, microbialites, and sediments. The MAM60+GFmix model fits the data much better and agrees with analyses of compositionally homogenized datasets with conventional site-heterogenous models. The consilience of evidence thus suggests that mitochondria is sister to the *Alphaproteobacteria* to the exclusion of MarineProteo1 and *Magnetococcia*. We also show that the ancestral presence of a crista-developing MICOS complex (a Mitofilin domain-containing Mic60) supports this relationship.

Introduction

Mitochondria stem from an ancient endosymbiosis that occurred during the origin of eukaryotic cells[1]. As a result, all extant eukaryotes have mitochondria or evolved from mitochondrion-bearing ancestors[1–3]. Some hypotheses have it that mitochondria provided excess energy required for the origin of eukaryotic complexity[4], whereas others suggest that mitochondrial symbiosis brought efficient aerobic respiration into a more complex proto-eukaryote[5]. The nucleocytoplasm of eukaryotes is now known to be most closely related to Asgard archaea[6–8]. Mitochondria, on the other hand, have been known for decades to be phylogenetically associated with the *Alphaproteobacteria*[9,10,1]. However, the precise relationship between mitochondria and the *Alphaproteobacteria*, or any of its sub-groups, has been elusive and remains a matter of intense debate (e.g., see [11,12]). Settling this debate will provide insights into the nature of the mitochondrial ancestor and the ecological setting of its endosymbiosis with the host cell[1].

Mitochondria have been placed in various regions of the tree of the *Alphaproteobacteria*. Most early studies suggested that mitochondria were most closely related to the *Rickettsiales*[13–20] (*Rickettsiales*-sister hypothesis), a group classically known for comprising intracellular parasites. This led many to believe that mitochondria evolved from parasitic alphaproteobacteria[18,21]. However, relationships between mitochondria and the *Pelagibacterales*[22,23], *Rhizobiales*[24], or *Rhodospirillales*[25] have also been proposed. These alternative proposals suggested that mitochondria may have evolved from either streamlined or metabolically versatile free-living alphaproteobacteria[22–25]. Most recently, the phylogenetic placement of mitochondria has been vividly debated[11,12]. One study found mitochondria as a sister group to the entire *Alphaproteobacteria* (i.e., the *Alphaproteobacteria*-sister hypothesis)[11]. This conclusion was supported by the inclusion of novel alphaproteobacterial metagenome-assembled genomes (MAGs) from worldwide

oceans, and by decreasing compositional heterogeneity through site removal. However, a subsequent study argued that removing compositionally heterogeneous sites from alignments might lead to the loss of true historical signal[26,12].The authors of the latter study, instead, used a taxon-removal and -replacement approach, and concluded that mitochondria branch within the *Alphaprotoebacteria* as sister to the *Rickettsiales* and some environmental metagenome-assembled genomes (MAGs)[12].

There are several reasons why it is difficult to confidently place mitochondria among their alphaproteobacterial relatives. First, the evolutionary divergence between mitochondria and their closest bacterial relatives is estimated to have occurred >1.5 billion years ago[27,28]. This has erased the historical signal (e.g., through multiple amino acid replacements) that was originally present in the few genes that mitochondria and alphaproteobacteria still share. Second, the *Alphaproteobacteria* is under sampled and most of its diversity remains to be discovered, as suggested by recent metagenomic surveys[11]. Third, and perhaps most problematic, the genomes of some lineages in the *Alphaproteobacteria* and those of mitochondria have undergone convergent evolution. For example, the *Rickettsiales* and *Holosporaceae* (intracellular bacteria)[29], or the *Pelagibacterales* and '*Puniceispirillaceae*' (planktonic bacteria)[30], have reduced or streamlined genomes with compositionally biased genes similar to those of mitochondria. The genes and genomes of these taxa are biased towards A+T nucleotides (and their proteins towards F, I, M, N, K, and Y amino acids) in contrast to other groups that have not evolved reductively (which might be biased towards G+C nucleotides and G, A, R, and P amino acids)[29]. This sort of compositional heterogeneity is often the cause of artefactual attractions among lineages with similar compositional biases in phylogenetic inference[31].

To cope with the aforementioned sources of phylogenetic errors, we developed and implemented a new phylogenetic model of protein evolution that accounts for compositional heterogeneity across both alignment sites and tree branches. Moreover, we also gathered an expanded set of 108 proteins of alphaproteobacterial origin in eukaryotes (in comparison to <67 previously available) and assembled more than 150 non-marine alphaproteobacterial MAGs from microbial mat, microbialite, and lake sediment metagenomes. We combined these improvements to explore and dissect the phylogenetic signal for the origin of mitochondria present in both modern eukaryotes and alphaproteobacteria.

Results

To date, most studies that aimed to phylogenetically place the mitochondrial lineage have relied exclusively on mitochondrion-encoded protein datasets that range from 12 to 38 proteins[16–18,32,11,12]. These markers are not only few (e.g., 24 genes and 6,649 sites in [11]) but tend to be compositionally biased because most mitochondrial genomes are rich in A+T. The only set of nucleus-encoded proteins of mitochondrial origin published thus far comprises 29 proteins[19,20].

To expand the number of proteins for placing the mitochondrial lineage, we systematically surveyed both nuclear and mitochondrial proteomes. After a multi-step phylogenetic screening, we identified 108 marker proteins of alphaproteobacterial origin in eukaryotes. Of these, 64 are exclusively nucleus-encoded, 27 are both nucleus- and mitochondrion-encoded, and 17 are exclusively mitochondrion-encoded proteins (Fig. 1A, Fig. S1). Our expanded dataset comprises most marker proteins previously identified[11,19,20] and adds 56 new ones (Fig. S1). Functional annotations show that these proteins have diverse functions within mitochondria (Fig. 1B, Table S1). Most are involved in energy metabolism (e.g., respiratory chain complex subunits) and protein synthesis (e.g., ribosomal subunits) (Fig. 1B, Table S1). The fact that all these proteins have mitochondrial functions strengthens the view that the genes that encode them were transferred from (proto-)mitochondria to nuclear genomes and are therefore not secondary lateral transfers to eukaryotes. The new nucleus-encoded proteins also tend to have much less variable and biased amino acid compositions in comparison to those which are mitochondrion-encoded and some that are both nucleus- and mitochondrion-encoded (Fig. 1A). Similarly, nucleus-encoded proteins also have a broader range of G A R P/F I M N K Y amino acid ratios of 0.70–1.95, whereas mitochondrion-encoded proteins have a range of 0.25–0.77 which suggests that they are much more compositionally biased towards F I M N K Y amino acids (and their genes towards A+T). The expanded set of nucleus-encoded

genes are expected to increase phylogenetic signal by virtue of increasing the amount of data, and also introduce potentially less compositionally biased sequences that could otherwise cause phylogenetic artefacts.

Most studies have exclusively relied on genomes of cultured alphaproteobacteria (e.g., [18–20,32]). Only one recent study incorporated novel alphaproteobacterial MAGs from metagenomes sequenced by the Tara Oceans project[11]. So far, all of these alphaproteobacterial MAGs came from oceanic open waters and tend to be small and A+T-rich[11]. Moreover, none of them appeared to be most closely related to mitochondria to the exclusion of other alphaproteobacteria[11].

To further increase taxonomic sampling across the *Alphaproteobacteria*, we assembled MAGs from metagenomes sequenced from diverse microbial mats, microbialites, and lake sediments (see Table S2 for details). In addition, we also screened MAG collections released previously[11,33–39], as well as the GTDB r89 database[40], for potentially phylogenetically novel alphaproteobacteria—together, these databases comprise more than ~ 3,300 alphaproteobacterial genomes and MAGs. The newly assembled MAGs were considerably diverse and widely distributed across the tree of the *Alphaproteobacteria* (Fig. 1C). Despite considerably expanding the sampled diversity of the *Alphaproteobacteria*, however, most of these new MAGs appear to fall within previously sampled major clades (Fig. 1C, Fig. 1D, Table S3), including those recently reported[11,40] (Fig. 1D, Table S3). The most novel MAGs include new members of the 'early-diverging' MarineProteo1 clade whose genomes are relatively small (1.43–2.71 Mbp) and not heavily compositionally biased towards A+T (43.6–59.7%) (Fig. S2, Table S4). In addition, several novel MAGs for 'basal' members of the *Rickettsiales* were found to be larger (1.47–2.36 Mbp) and enriched in G+C (49.2–61.2 or ~49.3% on average) relative to previously sampled members of this group (>0.6–2.11 Mbp and 32.1–34.2% G+C on average in the *Rickettsiaceae*, *Anaplasmataceae*, and *Midichloriaceae*) (Fig. S2, Table S4). The new alphaproteobacterial MAGs have moderate-to-high quality (according to criteria by [39,40]; 53.41–100% completeness and 0–9.17 redundancy), a wide range of G+C content (30.3–73.5%) and sizes (0.88–4.85 Mbp), and varying degrees of phylogenetic novelty (0.99–0.56 Relative Evolutionary Divergence score[40]) (Fig. 1D, Table S3)—this suggests that the methods used here to recover MAGs were not biased toward those with certain features (e.g., small sizes or high A+T content). Most of the new MAGs, which are widely distributed across the *Alphaproteobacteria* tree, also appear to encode an almost-complete set of bacteriochlorophyll biosynthesis enzymes which suggest that they come from photosynthesizers in the diverse environments sampled (e.g., microbial mats; Fig. 1D, Table S3).

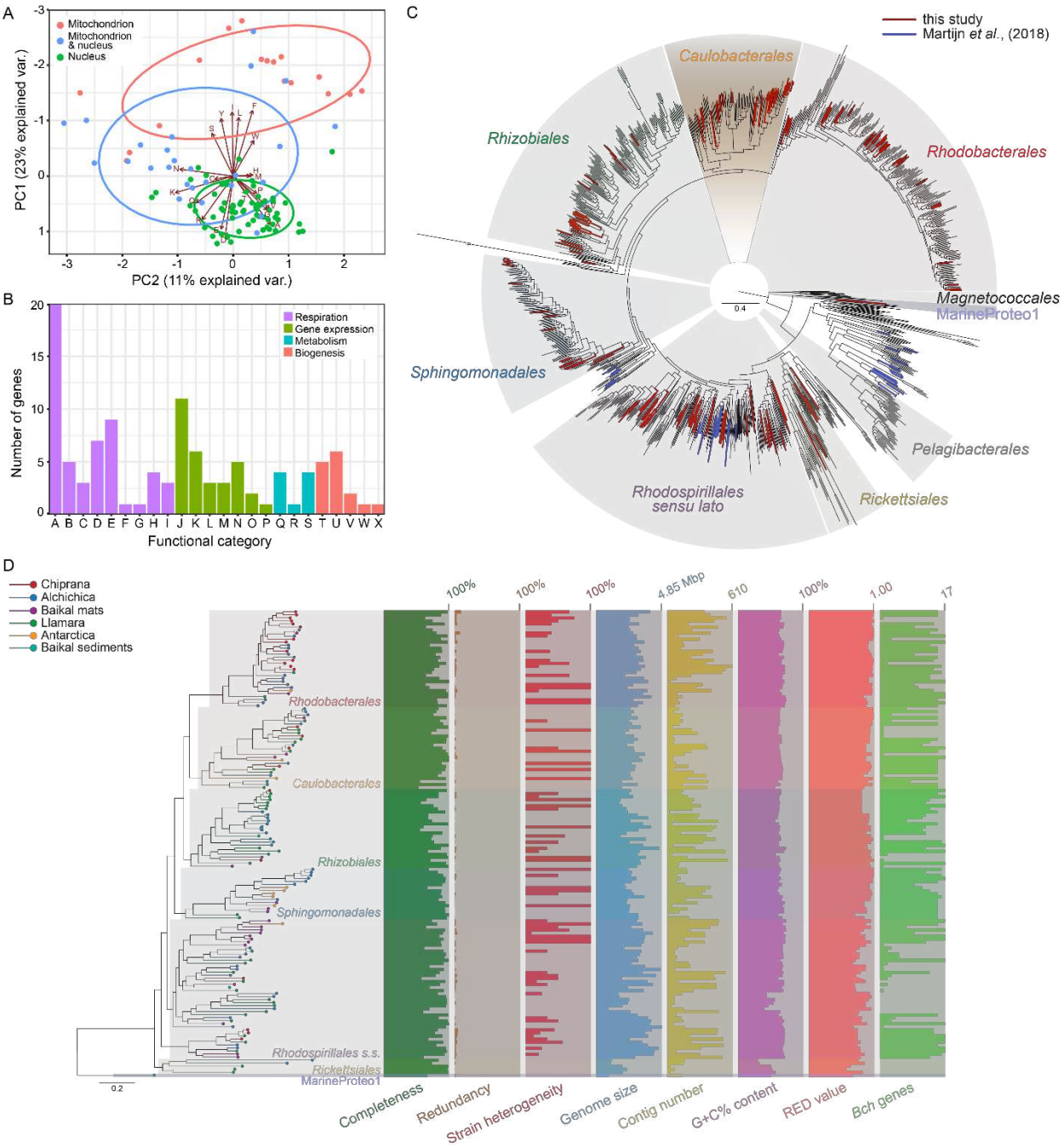


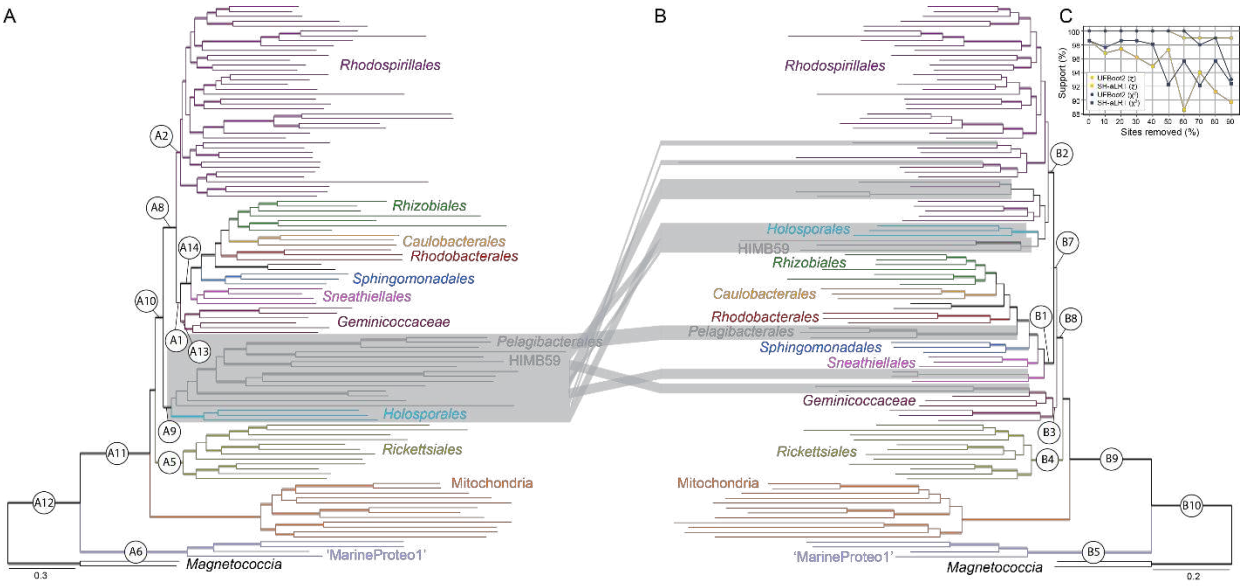
Figure 1. An expanded gene set and novel alphaproteobacterial MAGs from diverse environments. (**A**) Principal Component Analysis (PCA) of amino acid compositions for each one of the 108 mitochondrial genes of alphaproteobacterial origin used in this study. Mitochondrion-encoded genes (light red); Mitochondrion- and nucleus-encoded genes (light blue); nucleus-encoded genes (green); 95% confidence ellipses follow the same color code as genes. This PCA was inferred from alignments that contain only eukaryotes. (**B**) Functional classification of the marker genes of alphaproteobacterial origin in eukaryotes used for multi-gene phylogenetic analyses in this study. All these functions take place inside mitochondria. A: Complex I subunit/assembly factor; B: Complex II subunit/assembly factor; C: Complex III subunit/assembly factor; D: Complex IV subunit/assembly factor; E: Complex V subunit/assembly factor; F: Cytochrome c biogenesis; G: D-lactate dehydrogenase (respiratory chain); H: Pyruvate

dehydrogenase complex subunit; I: Krebs cycle; J: Ribosome large subunit; K: Ribosome small subunit; L: Ribosome translational factor; M: rRNA modification/maturation; N: tRNA modification/maturation; O: Aminoacyl-tRNA synthetase; P: RNA polymerase; Q: Branched-chain amino acid/fatty acid metabolism, R: Pyrimidine biosynthesis; S: Ubiquinone biosynthesis; T: Protein import/export; U: Iron-sulfur cluster biogenesis; V: Clp protease complex subunit; W: Proteasome-like complex subunit; X: Mitochondrial division (see also Table S1). (**C**) Phylogenetic tree of 154 novel MAGs reported here, the 45 MAGs reported by Martijn *et al*. (2018), and 1,188 of maximally diverse alphaproteobacterial genomes in GTDB r89 database. Taxon sample reduction was done with Treemmer[41] and phylogenetic inference with IQ-TREE (-fast mode) and the LG4X model (120 GTDB-Tk marker genes; 14,048 amino acid sites). (**D**). Phylogenetic tree for the 154 alphaproteobacterial MAGs reconstructed from diverse metagenomes sequenced in this study and summary of major features for each MAG. Tree was inferred with IQ-TREE (-fast mode) and the LG4X model after having removed 50% of most compositionally heterogeneous ζ sites (120 GTDB-Tk marker genes; 7,024 amino acid sites) (see also Table S3).

To address recent controversies[11,26,12], we first assembled the largest dataset to date that includes a new set of 64 nucleus-encoded and 44 mitochondrion-encoded proteins (108 genes in total and 33,704 amino acid sites; see above). Our dataset also comprised a wide taxon sampling with twelve mitochondria from diverse eukaryotes (from most 'supergroups'), and a broad set of 104 alphaproteobacteria that covered all major known lineages and maximized phylogenetic diversity (subsampled from a set of more than 3,300 genomes; see Methods). Importantly, our dataset incorporated several *Rickettsiales* species that have short branches and are less compositionally biased (Fig. 1D, Fig. S2, Table S4), as well as novel representatives of the MarineProteo1 clade (Fig. 1D, Fig. 2A, Table S4). Instead of relying on *Beta-*, and *Gammaproteobacteria* as outgroups (as in [11,12]), we used the much closer *Magnetococcia* which has been consistently found to be sister to all other alphaproteobacteria (e.g., [11,12,20]). This was done to decrease potential artefactual attractions between the long mitochondrial branch and distant outgroups, a concern raised before[11,26,12]. Furthermore, we also removed sites estimated to have undergone functional divergence at the origin of mitochondria (these represented only 5.2% of all sites) using the FunDi mixture model[42]. This was done to reduce potential artefacts from model misspecification as no phylogenetic model currently available adequately captures such patterns of functional divergence in proteins.

We first analyzed our dataset using the MAM60 site-heterogeneous model that was specifically inferred from our own dataset—this model has been shown to have a better fit than generic site-heterogenous models (e.g., C10-60)[43]. Analyses on the untreated dataset (i.e., without compositionally heterogeneous sites removed) placed mitochondria as sister to all of the *Alphaproteobacteria* with maximum support, i.e., both the monophyly of the *Alphaproteobacteria* and the *Alphaproteobacteria*-mitochondria clade were fully supported (Fig. 2A). However, these analyses also recovered the grouping between the *Pelagibacterales*, *Holosporaceae*, and other long-branching species (Fig. 2C, Mendeley Data) that, in previous work[29], were shown to artefactually attract each other because of similar amino acid compositional biases. A common strategy for dealing with compositional heterogeneity in the absence of site-and-branch-heterogeneous models is to remove alignment sites based on metrics that quantify their compositional heterogeneity[11,12,29]. The progressive removal of the compositionally most heterogeneous sites according to the ζ and χ^2 metrics[11,29,44] disrupted compositional attractions and showed clear support for the *Alphaproteobacteria*-sister hypothesis (Fig. 2B, Fig. 2C).

Because nucleus-encoded and mitochondrion-encoded proteins display different amino acid compositional patterns (Fig. 1A), we also analyzed these two protein sets separately. Whereas nucleus-encoded proteins unambiguously supported the *Alphaproteobacteria*-sister hypothesis across all analyses (Mendeley Data), the mitochondrion-encoded proteins showed decreased support for this hypothesis as compositionally heterogeneous sites are removed (Fig. S3, Mendeley Data). However, no alternative hypothesis was favored and any placement of mitochondria among the *Alphaproteobacteria* was unsupported for mitochondrion-encoded proteins (Fig. S3; Mendeley Data). This suggests that mitochondrion-encoded proteins may have a more equivocal phylogenetic signal. Unlike in many previous studies[19,20,12,11], we did not find support for the *Rickettsiales*-sister hypothesis in any of our analyses (Mendeley Data).



Tree / Model	B10	B5	A12	A6	B2	B1	B3	B7	B4	B8	B9	A2	A14	A13	A1	A9	A8	A5	A10	A11
MAM60 (no partitions)	1470 (3003129)	1469 (3003127)	842 (3002501)	841 (3002499)	792 (3002450)	768 (3002426)	743 (3002401)	743 (3002401)	653 (3002311)	649 (3002307)	626 (3002284)	182 (3001840)	176 (3001834)	166 (3001824)	159 (3001817)	89 (3001747)	63 (3001721)	33 (3001691)	25 (3001684)	0* (3001658)
Tree / Model	**A12**	**A6**	**B10**	**B5**	**A2**	**A14**	**A13**	**A1**	**A9**	**A8**	**B2**	**A5**	**A10**	**B1**	**A11**	**B7**	**B3**	**B4**	**B8**	**B9**
MAM60 (10 partitions)	1136 (-2990187)	1132 (-2990182)	834 (-2989884)	830 (-2989880)	495 (-2989545)	477 (-2989528)	446 (-2989497)	444 (-2989494)	339 (-2989389)	331 (-2989381)	294 (-2989344)	233 (-2989283)	220 (-2989271)	214 (-2989265)	207 (-2989257)	119 (-2989170)	87 (-2989138)	50 (-2989100)	44 (-2989094)	0* (-2989050)
Tree / Model	**B10**	**B5**	**A12**	**A6**	**B2**	**B1**	**B7**	**B3**	**B4**	**B8**	**B9**	**A2**	**A14**	**A13**	**A1**	**A9**	**A8**	**A5**	**A10**	**A11**
MAM60+GFmix (no partitions)	956 (-2983168)	954 (-2983166)	830 (-2983042)	828 (-2983040)	315 (-2982527)	306 (-2982518)	303 (-2982515)	293 (-2982505)	229 (-2982441)	215 (-2982426)	166 (-2982378)	161 (-2982373)	154 (-2982366)	141 (-2982353)	131 (-2982343)	120 (-2982332)	114 (-2982326)	32 (-2982244)	27 (-2982239)	0* (-2982212)
Tree / Model	**A12**	**A6**	**B10**	**B5**	**A2**	**A14**	**A13**	**A1**	**A9**	**A8**	**A5**	**A10**	**A11**	**B2**	**B1**	**B7**	**B3**	**B4**	**B8**	**B9**
MAM60+GFmix (10 partitions)	1382 (-2965660)	1381 (-2965659)	836 (-2965113)	830 (-2965108)	811 (-2965089)	785 (-2965063)	780 (-2965058)	779 (-2965057)	697 (-2964974)	685 (-2964962)	611 (-2964888)	605 (-2964882)	559 (-2964836)	216 (-2964494)	196 (-2964474)	161 (-2964438)	160 (-2964437)	51 (-2964328)	46 (-2964323)	0* (-2964278)

lnL

lnL

Figure 2. Phylogenetic tree of the *Alphaproteobacteria* and mitochondria, and support from our new site-and-branch-heterogeneous model. (**A**) Phylogenetic tree for the *Alphaproteobacteria* and mitochondria derived from a site-heterogeneous analyses of an untreated dataset. (**B**) Phylogenetic tree for the *Alphaproteobacteria* and mitochondria derived from a site-heterogeneous analysis of a dataset from which 50% of the most compositionally heterogeneous sites according to the ζ metric had been removed. The removal of this amount of ζ sites minimizes the variation of G A R P/F I M N K Y amino acid ratios across taxa (Table S5). The taxonomic labels follow the higher-level taxonomy outlined in [29]. Thickened branches represent branch support values of >90% SH-aLRT and >90% UFBoot2+NNI. (**C**) Variation in support values for the placement of mitochondria outside of the *Alphaproteobacteria* (SH-aLRT and UFBoot2+NNI) throughout the progressive removal of compositionally heterogenous sites according to the ζ and χ^2 metrics. Support for the branch that groups mitochondria with all alphaproteobacteria (but excludes MarineProteo1 and the *Magnetococcia*) is always maximum (i.e., 100% SH-aLRT /100% UFBoot2+NNI; Mendeley Data). (**D**) Heatmap table summarizing the differences in log-likelihoods (lnL) relative to the highest log-likelihood for several alterative placements of mitochondria (A1-14 and B1-B12 in (**A**) and (**B**); see Table S6 and Fig. S4 for all tree topologies) under a conventional site-heterogeneous model (MAM60) and our new site-and-branch-heterogeneous model (MAM60+GFmix). Models (rows) are arranged in increasing order (from top to bottom) according to lnL values. For each model (row), tree topologies (columns) are arranged in increasing order (from left to right) according to lnL values. Absolute log-likelihood values for each tree (A-T) under the different models tested are reported within parentheses. For all four models, all topologies other than the maximum-likelihood tree were rejected with *p*-values of < 0.0001 according to Bonferroni-corrected χ^2 tests. See Table S6 for all tree topologies and datasets tested.

All studies to date have exclusively relied on either site-homogenous or purely site-heterogeneous models (e.g., CAT in PhyloBayes or C60 in IQ-TREE)[11,12,14–20,22,23,32]. Indeed, no tractable model that accounts for compositional heterogeneity across branches and sites simultaneously is available; current branch-heterogeneous models cannot be combined with site-heterogenous models[31], or are too computationally intensive and suffer from convergence problems[45,46]. To overcome these shortcomings, we developed a model that captures the most important compositional heterogeneity in alphaproteobacterial genomes— namely the variation in the G A R P/F I M N K Y amino acid ratio that is driven by variation in G+C vs. A+T nucleotide content (see [29]). Our new branch-heterogeneous model, GFmix, models the variation in the ratio of G A R P/F I M N K Y amino acid frequencies across the phylogenetic tree in combination with conventional site-heterogeneous models (e.g., C10-60, MAM and UDM models). Briefly, this model requires a rooted tree, and introduces a new parameter that represents the G A R P/F I M N K Y ratio for every branch in a tree that is based on the amino acid compositions of all taxa that descend from that branch (see Materials and Methods for details). These parameters, in turn, adjust the frequencies of each site class in the site-profile mixture model resulting in a new transition rate matrix, Q_c, for each mixture class for the given branch. We developed and implemented the new GFmix model in a maximum likelihood framework.

To further test the phylogenetic placement of mitochondria, we used the MAM60+GFmix model to estimate log-likelihoods on two sets of fixed trees (Fig. 2A, Fig. 2B, Fig. S4). The first tree set was inferred from the untreated dataset (108 genes, 33,704 sites), whereas the second tree set was inferred from a compositionally homogenized dataset through site removal (108 genes, 16,029 sites); the latter dataset minimized the differences of G A R P/F I M N K Y amino acid ratios among taxa (Table S5). (Both tree sets were inferred using the MAM60 site-heterogeneous model; see above.) We then varied the position of mitochondria along all backbone branches on each fixed tree (Fig. 2A, Fig. 2B, Fig. S4). Furthermore, we also grouped proteins into partitions according to distances calculated based on their G A R P/F I M N K Y compositional disparity (Fig. S5). Our analyses show that likelihoods estimated under the MAM60+GFmix model improved significantly when compared to conventional site-heterogeneous models (Fig. 2D, Table S6, likelihood ratio test (LRT) p-value = 0); model fit was improved even more when the proteins were grouped into ten separate partitions according to G A R P/F I M N K Y compositional disparity (Fig. 2D, Table S6, LRT p-value = 0). Importantly, the partitioned MAM60+GFmix model clearly favours trees that display the *Alphaproteobacteria*-sister relationship and where the grouping of long-branching and compositionally biased taxa (e.g., *Pelagibacterales*, *Holosporaceae*) is disrupted (i.e., those trees recovered from compositionally homogenized datasets through ζ site removal; Fig. 2D, Table S6). This suggests that the removal of ζ sites effectively decreases overall compositional heterogeneity and potential artefacts.

The top three trees often favored by the MAM60+GFmix model (i.e., those with the highest likelihoods) have mitochondria in adjacent branches: *Alphaproteobacteria*-sister (trees A11 and B9 in Fig. 2A and Fig. 2B), *Rickettsiales*-sister (trees A5 and B4 in Fig. 2A and Fig. 2B), and mitochondria as sister to all alphaproteobacteria except the *Rickettsiales* (or *Caulobacteridae*-sister; trees A10 and B8 in Fig. 2A and Fig. 2B)[29,47]. However, Bonferroni-corrected χ^2 topology tests show that the optimal trees that display the *Alphaproteobacteria*-sister relationship are significantly better than all trees with other positions for mitochondria (see Fig. 2D). Even though the *Alphaproteobacteria*-sister relationship is also favored by the MAM60+GFmix model for the mitochondrion-encoded protein dataset, the *Caulobacteridae*-sister relationship cannot be rejected by the Bonferroni-corrected χ^2 tests (i.e., p-values > 0.05; Table S6). This further supports the notion that the phylogenetic signal for the placement of mitochondria is weaker in mitochondrion-encoded proteins (see above). The *Rickettsiales*-sister relationship is rejected for all datasets and models (p-value < 0.005; Table S6). Overall, most of our distinct phylogenetic approaches show support for the *Alphaproteobacteria*-sister hypothesis.

Discussion

We have found significant support for the *Alphaproteobacteria*-sister hypothesis that has the mitochondrial lineage as the closest sister to all currently sampled alphaproteobacteria. Our findings thus conflict with the recent suggestion that mitochondria may branch within the *Alphaproteobacteria* as sister to the *Rickettsiales*[12]. Indeed, we believe that the design of the study by Fan *et al*., (2020) was particularly prone to artefacts. In an effort to choose less compositionally biased (i.e., G+C-rich) species for mitochondria and the *Rickettsiales*, these authors inadvertently selected species that are more divergent than most members of their respective groups. For example, the inclusion of mitochondria of flowering plants led to a considerably long stem branch for the mitochondrial lineage (see their Fig. S31-48). Similarly, *Anaplasma*, *Neorickettsia*, and *Wolbachia* (*Anaplasmataceae*) are among the longest branches in the *Rickettsiales* (see their Fig. S50; see also our Fig. S2). All these species are secondarily, and not ancestrally, less compositionally biased, i.e., they evolved from species with A+T-rich genomes. Moreover, their analyses were based on a rather small dataset that comprised only 18 or 24 mitochondrion-encoded genes (5,583 and 6,643 sites, respectively) and fewer than 41 taxa. These factors may, in combination, have led to the inference of poorly supported trees (e.g., see their Figs. S31-40), and an artefactual attraction between mitochondria, the *Rickettsiales*, and the FEMAG I and II groups (i.e., Fast-Evolving MAG I and II; see their Fig. 4).

Several previous studies have suggested that mitochondria were either sister to the *Rickettsiales*[18–20] or phylogenetically embedded in a larger group comprised of both the *Rickettsiales* and the *Holosporaceae*[20]. These hypotheses implied that the mitochondrial ancestor may have been an intracellular parasite: throughout its early evolution, the ancestor of mitochondria changed its function from an energy parasite to an ATP-producing respiratory organelle[18–21]. The finding that mitochondria are no longer phylogenetically associated to the *Rickettsiales* and are instead sister to the entire *Alphaproteobacteria* clade makes a parasitic origin of mitochondria less plausible. However, the nature of the mitochondrial ancestor remains poorly constrained. Future studies on species of the MarineProteo1 clade might shed some light on the early evolution of the *Alphaproteobacteria*, and possibly also on the mitochondrial ancestor. However, we note that the MarineProteo1 clade is separated by a long branch from the *Alphaproteobacteria* and mitochondria. Currently available genomes for the MarineProteo1 clade are relatively small, but not necessarily compositionally biased, and suggest that these alphaproteobacteria might be reduced and physiologically specialized (Fig. S2, Table S4).

Unravelling the deep evolutionary history of mitochondria is an inherently hard phylogenetic problem. One of the main challenges is to properly account for the drastically different compositional biases across anciently diversified lineages[29]. Here, we have moved towards overcoming this major obstacle. Our newly developed and implemented site-and-branch-heterogenous model allowed us, for the first time, to test different phylogenetic placements for mitochondria relative to the *Alphaproteobacteria* while accounting for the drastic amino acid compositional changes that alphaproteobacterial and mitochondrial proteins have undergone. A consilient view emerges from the combination of modelling and reducing compositional heterogeneity: the *Alphaproteobacteria*-sister hypothesis is robust and unlikely to be artefactual. However, we caution that the phylogenetic signal preserved in mitochondrion-encoded proteins is weak and ambiguous. The recovery of the *Rickettsiales*-sister relationship in previous studies[11,12] may thus be result of ambiguous phylogenetic signal and long-branch attraction. Therefore, we suggest that it is currently best to view mitochondria as an early offshoot of the alphaproteobacterial lineage that diverged just prior to the diversification of known extant groups. This is suggested by the short internal branch lengths between mitochondria and *Alphaproteobacteria* (see Fig. 2A, Fig. 2B) and is supported by the shared presence of the Mitochondrial Contact Site and Cristae Organizing System (i.e., a Mitofilin domain-containing Mic60) in only mitochondria and the *Alphaproteobacteria*, but not in members of the *Magnetococcia* and MarineProteo1[48,49] (Fig. S2, Table S4). Future efforts should focus on exploring diverse environments for unknown and extant alphaproteobacterial lineages that may be more closely related to mitochondria.

Materials and Methods

Metagenomic sequencing and MAG assembly

Samples collected from (1) microbial mats in the Salada de Chiprana (Spain, December 2013), Salar de Llamara[50], Lakes Bezymyannoe and Reid (Antarctica, January 2017) and several hot springs around Lake Baikal (Southern Siberia, July 2017), (2) microbialites in Lake Alchichica[51], and (3) sediments in Lake Baikal, were fixed in ethanol (>70%) *in situ* and stored at -20°C as previously described[50]. Total DNA was purified from samples using the DNeasy PowerBiofilm Kit (QIAGEN, Germany) by following the manufacturer's guidelines. DNA extracted from microbialite fragments was further cleaned using the DNeasy PowerClean Cleanup Kit (QIAGEN, Germany) as previously described[52]. DNA was quantified using Qubit®. DNA library preparation and sequencing were performed with an Illumina HiSeq2000 v3 (2x100 bp paired-end reads) by Beckman Coulter Genomics (Danvers, MA, USA), and with an Illumina HiSeq2500 (2x125 bp paired-end reads) by Eurofins Genomics (Ebersberg, Germany). A summary of the metagenomic libraries sequenced can be found in Table S2.

Raw Illumina short reads from all sequenced Illumina paired-end libraries were quality-assessed with FastQC v.0.11.7 and quality-filtered with Trimmomatic v.0.36[53]. Libraries made from samples from Lake Alchichica and the Llamara saltern were processed with the following workflow. Libraries were individually assembled, and technical replicates co-assembled (Table S2), with metaSPAdes v.3.10.0[54]. Contigs smaller than 2,500 bp in the (co-)assemblies were removed. Filtered reads were then individually mapped onto each assembly with Bowtie2 to obtain contig coverages[55]. Contigs were binned using MaxBin v.2.2.2 which relies on differential coverage across samples, tetranucleotide composition and single-copy marker genes[56]. The completeness and contamination of the bins reported by MaxBin v.2.2.2 were assessed with CheckM v.1.0.12[57]. Genome bins that were phylogenetically affiliated to the *Alphaproteobacteria* based on the manual examination of the CheckM reference genome tree (itself based on the concatenation of 43 marker genes) were retained. Reads were then individually mapped onto each alphaproteobacterial genome bin with Bowtie2. All paired and unpaired reads that successfully mapped to the alphaproteobacterial bins were subsequently co-assembled with metaSPAdes. The resulting co-assembly was processed through the Anvi'o metagenomic workflow[58]. In brief, reads were mapped to the final metaSPAdes co-assembly with Bowtie2 to obtain contig coverage values. DIAMOND searches[59] of predicted proteins against the NCBI GenBank nr database were done to assign taxonomic affiliations to each contig. CONCOCT2[60], implemented in the Anvi'o suite, was used to bin the resulting metagenome. Contigs were organized according to the composition and coverage by anvi-interactive. The predicted CONCOCT2 bins were visualized and manually refined based on their composition, coverage, taxonomy and completeness/redundancy. Libraries made from samples from Antarctica, the Chiprana saltern and Lake Baikal were processed with the following workflow. Libraries from the same location or environment type were co-assembled with MEGAHIT v.1.1.1[61]. Contigs smaller than 2,500 bp in the co-assemblies were removed. Filtered reads were then individually mapped onto each co-assembly with Bowtie2 to obtain contig coverages. Contigs were binned using three different binners (MetaBAT v.2.12.1[62], MaxBin 2.2.4[56], CONCOCT2[60]) and their results were combined into consensus contigs bins with DAS Tool v.1.1.0[63].

Marker protein selection

We built an expanded dataset of mitochondrion- and nucleus-encoded proteins of alphaproteobacterial origin in eukaryotes. For the nucleus-encoded proteins, BLAST[64] similarity searches of all proteins contained in the predicted proteomes of 13 representative eukaryotes were conducted against a database of 176 prokaryotes (136 bacteria and 40 archaea). BLAST hits were clustered into homologous families with a custom Perl script, aligned with MAFFT and the L-INS-I method[65], and then trimmed with BMGE[66]. Phylogenetic trees for each homologous gene family were inferred under the LG model in RAxML v.8[67]. These trees were then sorted based on the criterion that eukaryotes form a clade with alphaproteobacteria. Manual inspection of the trees then followed to remove paralogs and contaminants. For mitochondrion-encoded genes, mitochondrial clusters of orthologous genes (MitoCOGs)[68] that are widespread among eukaryotes were used.

Both mitochondrion-, and nucleus-encoded candidate marker proteins were then compared through BLAST searches against those reported previously by Wang and Wu (2015)[20] and Martijn *et al*., (2018)[11]. Our dataset encompassed most proteins from these other datasets, with few exceptions. The non-redundant and remaining candidate marker proteins comprising the union of these five datasets, were then further screened phylogenetically. Using a representative eukaryotic (mitochondrial) query for each marker gene, BLAST searches were done against a database that comprises 107 diverse bacteria (representing 27 cultured phyla) and 23 diverse eukaryotes (representing 6 major groups); eukaryotes were selected based on the availability of both mitochondrial and nuclear genomes or transcriptomes (see Table S7). Homologues were aligned with MAFFT, alignments trimmed with TrimAl[69] and single-protein trees inferred with IQ-TREE[70]. The single-protein trees were inspected visually to remove duplicates, paralogues, and any other visual outlier such as extremely divergent sequences. Single-protein trees were then re-inferred from the curated alignments and visually inspected. Proteins for which trees showed a sister relationship between eukaryotes and alphaproteobacteria were kept for further analyses. Finally, these candidate marker proteins were annotated and further refined using the EggNOG database and BLASTp searches. The final marker proteins set comprised 108 genes, 64 of which are exclusively nucleus-encoded, 17 are exclusively mitochondrion-encoded, and 27 are both mitochondrion- and nucleus-encoded (Fig. S1). The annotations confirm that all marker proteins are predicted to be localized to mitochondria in eukaryotes (Table S1).

Dataset assembly

To increase taxon sampling as much as possible, MAGs reported in Anantharaman *et al*., (2016)[33], Graham *et al*., (2018)[34], Delmont *et al*., (2018)[35], Martijn *et al*., (2018)[11], Mehrshad *et al*., (2016)[36], Tully *et al*., (2017)[37], Tully *et al*., (2018)[38] and Parks *et al*., (2017)[39] were added to those reconstructed here (see Metagenomic analyses). To improve the quality of our MAG selection, MAGs were analyzed with the CheckM lineage workflow and those with quality values (completeness – 5x contamination) lower than 50 were discarded, just as done before by Parks *et al*., (2017, 2018)[39,40]. MAGs were then filtered according to their taxonomic affiliation to the *Alphaproteobacteria*. A phylogenetic tree for all MAGs and all *Proteobacteria* taxa in the GTDB r89 database[40] was inferred from 120 marker proteins, built-in in the GTDB-Tk software, using IQ-TREE v.1.6.10[70] and the LG4X+F model. To increase phylogenetic accuracy, a second tree was inferred with the LG+PMSF(C60)+G4+F using the LG4X tree as guide. All MAGs that fell within the *Alphaproteobacteria* clade in the GTDB-Tk tree were chosen for subsequent analyses. Together, these added up to more than 3,300 alphaproteobacteria. In order to reduce computational burden, Treemmer v.0.1b was then used to reduce the number of alphaproteobacterial taxa from the GTDB-TK tree while maximizing phylogenetic diversity[41]. The Treemmer analysis was constrained so representatives from major clades, as visually identified, were retained. Finally, a set of reference alphaproteobacteria (formally described species) were added, and long-branching alphaproteobacteria were replaced by short-branching relatives.

To retrieve homologues, PSI-BLAST searches with either one, two, or three iterations using representative mitochondrial (eukaryotic) query sequences for each marker protein were done against a database that comprised all carefully selected predicted proteomes. To remove non-orthologous sequences, homologous protein sets were retrieved for each marker protein, aligned with MAFFT, trimmed with TrimAl and trees inferred with IQ-TREE. The single-protein trees were visually inspected to remove duplicates, paralogues, and any other visual outlier such as extremely divergent sequences. The curated homologous protein sets were finally aligned again with MAFFT v.7.3.10 and the L-INS-I method. To increase phylogenetic signal by removing poorly aligned and non-homologous aligned regions, Divvier v.1.0 was used with the -partial and -mincol options[71]. Only sites with more than 10% of data were retained. To reduce incongruency among proteins due to, for example, lateral gene transfer, Phylo-MCOA v.1.4[72] was employed on single-protein trees with UFBoot2+NNI as branch support which were inferred with IQ-TREE v.1.6.10 and the best-fitting model as identified by Model-Finder[70,73]. Single-protein alignments were concatenated with SequenceMatrix v.1.8[74].

Phylogenetic analyses using site-heterogeneous models

For multi-protein phylogenetic analyses on the supermatrix, trees were first inferred in IQ-TREE v.1.6.10 under the LG4X+F model. The resulting site-homogenous tree was then used as a guide tree to infer a new phylogenetic tree under the LG+PMSF(C60)+F+G4 model[75]. Consequently, the resulting site-heterogenous tree was used as a guide tree to infer a new phylogenetic tree under the dataset-specific LG+PMSF(MAM60)+F+G4 model. The dataset-specific MAM60 model was estimated using the MAMMaL software[43]. This site-heterogeneous mixture model is directly inferred from the dataset analyzed and therefore is more specific than the general C10-60 mixture models. To account for more than 60 (e.g., C60 or MAM60) amino-acid composition profiles across the data, we used the general UDM128 mixture model as LG+UDM128+G4+F that allows for 128 amino acid composition profiles[76]. The software FunDi was used to estimate functionally divergent sites in the branch that separates the mitochondrial lineage from all other taxa[42]. Sites with a probability of being functionally divergent > 0.5 were removed. Progressive removal of compositionally heterogeneous sites was performed according to the ζ and the χ^2 metrics/methods as described before[11,29,44]. Both metrics are designed to estimate compositional heterogeneity per site based on different criteria.

Bayesian analyses were conducted with PhyloBayes MPI v1.8 using the CAT-GTR+G4 model[77,78]. PhyloBayes MCMC chains were run for >20,000 cycles or until convergence between the chains was achieved and the largest discrepancy in posterior probabilities for splits between chains ('max-diff') was <0.1. Individual chains were summarized into a Bayesian consensus tree using a burn-in of 500 trees and subsampling every 10 trees. However, most chains did not reach convergence or resolve the phylogenetic placement of mitochondria relative to alphaproteobacterial lineages (Mendeley Data).

Phylogenetic analyses using the site-and-branch-heterogeneous GFmix model

The site profile mixture models discussed above have C site frequency profiles and a K-class discretized gamma mixture model for site rates. Under these models, the likelihood of site pattern $\mathbf{x_i}$ at site *i* is given by:

$$P(\mathbf{x_i}; w_c, \boldsymbol{\theta}) = \sum_{\boldsymbol{c=1}}^{\boldsymbol{C}} w_c \sum_{k=1}^{K} P\left(\mathbf{x_i} \mid r_k, \boldsymbol{\pi}^{(c)}; \boldsymbol{\theta}\right)/K$$

Where r_k is the site rate of gamma-rates class *k*, $\boldsymbol{\pi}^{(c)}$ is the vector of amino acid frequencies in class *c* of the site-profile mixture model, w_c is the class weight and θ is the vector of other adjustable parameters (branch lengths, α shape parameter and tree topology) in the model. In order to model shifts in the relative frequencies of the amino acids G A R P (specified by G+C-rich codons) and F I M N K Y (specified by A+T-rich codons) in different branches of the tree, the foregoing vectors of amino acid frequencies, $\boldsymbol{\pi}^{(c)}$, are modified in a branch-specific manner in the following way.

Let *b* denote the ratio of aggregate frequencies of G A R P to F I M N K Y amino acids; i.e., $b := \pi_G/\pi_F$ for $\pi_G = \sum_{j \in \{G,A,R,P\}} \pi_j$ and $\pi_F = \sum_{j \in \{F,Y,M,I,N,K\}} \pi_j$ where π_j is the frequency of amino acid *j*. For every branch *e* in the phylogenetic tree under consideration, we can obtain estimates by a hierarchical procedure where b_e is obtained from the GARP/FIMNKY ratio of all of the sequences at the tips of the tree that descend from branch *e*. Using these estimates, the values in the class frequency vectors, $\boldsymbol{\pi}^{(c)}$, for any site profile class are modified in the following way to be branch-*e*-specific class frequencies, $\boldsymbol{\pi}^{(ce)}$. The modified class frequencies have to satisfy a number of constraints including:

$$\pi_j^{(ce)} = \begin{cases} \mu^{(ce)} S_G^{(e)} \pi_i^{(c)} & j \in \{G, A, R, P\} \\ \mu^{(ce)} S_F^{(e)} \pi_j^{(c)} & j \in \{F, Y, M, I, N, K\} \\ \mu^{(ce)} \pi_j^{(c)} & otherwise \end{cases}$$

and $\sum_j \pi_j^{(ce)} = 1$ and

449 $$\frac{\sum_{c=1,j\in\{G,A,R,P\}}^{C} w_c \pi_j^{(ce)}}{\sum_{c=1,j\in\{F,Y,M,I,N,K\}}^{C} w_c \pi_j^{(ce)}} = b_e$$

450 This leads to non-linear equations for $\mu^{(ce)}$, $S_G^{(e)}$ and $S_F^{(e)}$ that are solved numerically for each branch *e* to
451 generate the modified class frequencies. For each branch and site class *c*, $\pi_j^{(ce)}$ values are used to create
452 a new transition $Q^{(ce)}$ matrix for likelihood calculations for all site patterns over that branch. The same
453 approach is used with frequencies coming all extant taxa to obtain the root frequencies. A software
454 implementation of GFmix is available at https://www.mathstat.dal.ca/~tsusko/software.html.

455 Partitioning the data matrix for GFmix calculations.

456 The foregoing framework assumes that for each aligned protein in a given concatenated dataset, the
457 GARP/FIMNKY ratios (b_e's) for every branch in the tree will be similar. However, for our data matrix this
458 assumption is not true as different proteins show different degrees of GARP/FIMNKY variation across
459 taxa depending on the location of the corresponding gene (e.g., nucleus-encoded vs. mitochondrial-
460 encoded) and degree of conservation. For this reason, we clustered the proteins in our dataset into
461 groups in the following way. For each protein *v* and each taxon *t* we calculated the GARP/FIMINKY ratio,
462 $b_v^{(t)} = \pi_G^{(t)}/\pi_F^{(t)}$. Then, we calculated the overall distance between these ratios for every pair of proteins *u*
463 and *v* in the data matrix as $d_{u,v} = \sum_t |b_v^{(t)} - b_u^{(t)}|/N_{u,v}$ where $N_{u,v}$ is the total number of taxa for which
464 sequences were available for both proteins (this normalization accounts for the differing amounts of
465 missing data for different proteins). The proteins were then clustered based on $d_{u,v}$ distances using the
466 UPGMA algorithm in MEGA-X[79] and 10 clusters were chosen as a computationally tractable number of
467 partitions for further analysis. The GFmix model was then applied to these 10 partitions allowing for
468 separate b_e values and branch lengths for each partition. The overall log-likelihoods for topologies were
469 obtained as the sum of log-likelihoods of that topology over all partitions.

470 To test the relative fits of the foregoing phylogenetic models to the data we used likelihood ratio tests
471 (LRTs). Briefly, the log-likelihood of a given mixture model (e.g., MAM60) under its optimal tree was
472 compared to the log-likelihood of the corresponding mixture-GFmix model. The former model is a special
473 case of the latter where all the b_e parameters are equal to the overall GARP/FIMNKY ratio. The likelihood
474 ratio statistic LRS, which is defined as twice the difference in these log-likelihoods, was calculated and a
475 *p*-value was determined as $P[\chi_d^2 > LRS]$ where *d* is the difference the number of additional parameters in
476 the more complex model (i.e., the b_e parameters); here *d*=2t-2 where t is the number of taxa. A similar
477 approach is taken to compare the partitioned models to the non-partitioned models. In this case there
478 were additional branch lengths and b_e parameters for each partition and so for 10 partitions, *d*=9(2t-
479 2)+9(2t-3). We note that this test is conservative because b_e estimates were not determined by maximum
480 likelihood. Therefore, the true *p*-values for the LRTs are less than $P[\chi_d^2 > LRS]$. If the LRT rejects the null
481 hypothesis under these conditions, then the correct test would also reject.

482 Topology testing using the Bonferroni-corrected χ^2 test.

483 The topology test is a variation of the chi-squared test presented in Susko (2014)[80] that corrects for
484 selection bias. The chi-squared test is a test of two trees. The null hypothesis $H_0: \tau = \tau_0$ is tested against
485 $H_A: \tau = \tau_A$ where τ is the true topology. As a test statistic, it uses the likelihood ratio statistic, *LRS*, which
486 is defined as twice the difference between the maximized log likelihood when the true topology is τ_A and
487 the maximized log likelihood for τ_0. It gives a *p*-value $p(\tau_A) = P[\chi_d^2 > LRS]$, the probability that a chi-
488 squared random variable with *d* degrees of freedom is greater than the observed *LRS.* Here the degrees
489 of freedom, *d,* are determined as the number of branches that are 0 in the consensus tree representing
490 both τ_0 and τ_A.

491 In the absence of a particular τ_A of interest, to test whether $H_0: \tau = \tau_0$ can be rejected, we consider the
492 alternative $H_A: \tau = \hat{\tau}$, where $\hat{\tau}$ is the maximum likelihood (ML) topology. Because the topology under the

alternative hypothesis was selected based on the data rather than being fixed *a priori*, this can induce a selection bias[81]. The Bonferroni approach uses a input set of trees and approximates the *p*-value when $H_A: \tau = \hat{\tau}$ by the Bonferroni-corrected *p*-value one would obtain testing $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$, $i \in A$ where A is the set of input trees that are compatible with the consensus tree of τ_0 and $\hat{\tau}$.

The approximation is based on probability calculations treating the consensus tree of $\hat{\tau}$ and τ_0 as the true tree. This is consistent with what is done in the chi-square test and in testing more generally, where one often calculates *p*-values under parameters on the boundary between the null and alternative hypotheses spaces (see [80] for additional discussion). If the true tree is the consensus tree, then it is likely that the ML topology will be in A. Because the largest likelihood is the one corresponding to $\hat{\tau}$, the smallest *p*-value among the $n(A)$ *p*-values obtained by testing $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$ is likely to be $p(\hat{\tau})$; there is some possibility that a tree with a smaller degrees of freedom would give the smallest *p*-value, so this is an approximation. In summary, $p(\hat{\tau})$ is approximately the same as the minimum *p*-value obtained by testing $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$.

Rephrasing the test as approximately the same as the result of multiple tests $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$, $i \in A$ lays bare that multiple testing is the source of selection bias. Bonferroni correction is a widely used approach to adjusting for multiple testing. As one final approximation, rather than using the usual Bonferroni-corrected *p*-value, $\mathrm{n(A)\ p(\hat{\tau})}$, we use the exact correction had the *p*-values coming from the tests been independent,

$$1 - [1 - p(\hat{\tau})]^{n(A)}.$$

This *p*-value is approximately the same as the usual Bonferroni correction when $n(A)\ p(\hat{\tau})$ is small, which is the case of greatest interest, but has the advantage of always being between 0 and 1. Additional information about the Bonferroni correction is available in [82].

Profile Hidden Markov Model (pHMM) searches

To search for bacteriochlorophyll enzymes, a set of 17 custom-made pHMMs for the genes *bchB, bchC, bchD, bchE, bchF, bchG, bchH, bchI, bchJ, bchL, bchM, bchN, bchO, bchP, bchX, bchY, bchZ* was used against predicted proteomes from the MAGs reconstructed in this study. These pHMMs were created from manually curated sets of *bch* genes from diverse proteobacteria. The searches were done with the program hmmsearch of the HMMER suite using an E-value cut-off of 1E-25. To search for mitofilin-domain containing *mic60* genes, the Pfam pHMM for Mitofilin (PF09731) was used with its own GA cut-off value.

Data Availability

Sequencing data were deposited in NCBI GenBank under the BioProjects PRJNA315555, PRJNA438773, PRJNAXXXXXX, PRJNAXXXXXX, PRJNAXXXXXX, and PRJNA703749. Assembled metagenomes, novel alphaproteobacterial MAGs, and gene files (unaligned, aligned, and aligned and trimmed) are available at: DOI: 10.6084/m9.figshare.14355845. Datasets and phylogenetic trees inferred in this study are available at: DOI: http://dx.doi.org/10.17632/dnbdzmjjkp.1. The GFmix model software is available at: https://www.mathstat.dal.ca/~tsusko/software.html.

Acknowledgements

SAM-G is supported by a EMBO Postdoctoral Fellowship (ALTF 21-2020). We are thankful to Bruce Curtis (Dalhousie University) and Dayana Salas-Leiva (Dalhousie University) for assistance with scripts, and to Wendy Valencia (Harvard University) and Camilo Calderon (Rutgers University) for advice on Python and R. This work was supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grants 735923LPI (DOI: https://doi.org/10.46714/735923LPI) awarded to AJR and GBMF9739 (DOI: https://doi.org/10.37807/GBMF9739) awarded to PLG, and Discovery Grants from the Natural Sciences and Engineering Research Council of Canada awarded to AJR, ES, and CHS.

References

1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Current Biology* **27**, R1177–R1192 (2017).
2. Stairs, C. W., Leger, M. M. & Roger, A. J. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Phil. Trans. R. Soc. B* **370**, 20140326 (2015).
3. Müller, M. *et al.* Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiol. Mol. Biol. Rev.* **76**, 444–495 (2012).
4. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).
5. Cavalier-Smith, T. Predation and eukaryote cell origins: A coevolutionary perspective. *The International Journal of Biochemistry & Cell Biology* **41**, 307–322 (2009).
6. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
7. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
8. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).
9. Gray, M. W. Mitochondrial Evolution. *Cold Spring Harb Perspect Biol* **4**, a011403 (2012).
10. Gray, M. W. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *PNAS* **112**, 10133–10138 (2015).
11. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
12. Fan, L. *et al.* Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nature Ecology & Evolution* **4**, 1213–1219 (2020).
13. Viale, A. M. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Letters* **341**, 146–151 (1994).
14. Andersson, S. G. E. *et al.* The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
15. Wu, M. *et al.* Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biol.* **2**, E69 (2004).
16. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome Phylogenies Indicate a Meaningful A-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales. *Mol Biol Evol* **23**, 74–85 (2006).
17. Williams, K. P., Sobral, B. W. & Dickerman, A. W. A robust species tree for the alphaproteobacteria. *J. Bacteriol.* **189**, 4578–4586 (2007).
18. Sassera, D. *et al.* Phylogenomic Evidence for the Presence of a Flagellum and cbb3 Oxidase in the Free-Living Mitochondrial Ancestor. *Mol Biol Evol* **28**, 3285–3296 (2011).
19. Wang, Z. & Wu, M. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PLoS ONE* **9**, e110685 (2014).
20. Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Sci Rep* **5**, 7949 (2015).
21. Ball, S. G., Bhattacharya, D. & Weber, A. P. M. Pathogen to powerhouse. *Science* **351**, 659–660 (2016).
22. Thrash, J. C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Scientific Reports* **1**, 13 (2011).
23. Georgiades, K., Madoui, M.-A., Le, P., Robert, C. & Raoult, D. Phylogenomic Analysis of Odyssella thessalonicensis Fortifies the Common Origin of Rickettsiales, Pelagibacter ubique and Reclimonas americana Mitochondrion. *PLoS ONE* **6**, e24857 (2011).
24. Abhishek, A., Bavishi, A., Bavishi, A. & Choudhary, M. Bacterial genome chimaerism and the origin of mitochondria. *Can. J. Microbiol.* **57**, 49–61 (2011).
25. Thiergart, T., Landan, G., Schenk, M., Dagan, T. & Martin, W. F. An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin. *Genome Biol Evol* **4**, 466–485 (2012).
26. Gawryluk, R. M. R. Evolutionary Biology: A New Home for the Powerhouse? *Current Biology* **28**, R798–R800 (2018).

27. 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).
28. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
29. Muñoz-Gómez, S. A. *et al.* An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *eLife* **8**, e42535 (2019).
30. Luo, H. Evolutionary origin of a streamlined marine bacterioplankton lineage. *ISME J* **9**, 1423–1433 (2015).
31. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).
32. Rodríguez-Ezpeleta, N. & Embley, T. M. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS ONE* **7**, e30520 (2012).
33. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* **7**, 13219 (2016).
34. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *The ISME Journal* **12**, 1861–1866 (2018).
35. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology* **3**, 804–813 (2018).
36. Mehrshad, M., Amoozegar, M. A., Ghai, R., Shahzadeh Fazeli, S. A. & Rodriguez-Valera, F. Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the Brackish Waters of the Caspian Sea. *Appl Environ Microbiol* **82**, 1599–1612 (2016).
37. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**, e3558 (2017).
38. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, (2018).
39. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533–1542 (2017).
40. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996–1004 (2018).
41. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
42. Gaston, D., Susko, E. & Roger, A. J. A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics* **27**, 2655–2663 (2011).
43. Susko, E., Lincker, L. & Roger, A. J. Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. *Molecular Biology and Evolution* **35**, 1266–1283 (2018).
44. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
45. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).
46. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **25**, 842–858 (2008).
47. Ferla, M. P., Thrash, J. C., Giovannoni, S. J. & Patrick, W. M. New rRNA gene-based phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry and phylogenetic instability. *PLoS ONE* **8**, e83383 (2013).
48. Muñoz-Gómez, S. A. *et al.* Ancient Homology of the Mitochondrial Contact Site and Cristae Organizing System Points to an Endosymbiotic Origin of Mitochondrial Cristae. *Current Biology* **25**, 1489–1495 (2015).
49. Muñoz-Gómez, S. A., Wideman, J. G., Roger, A. J. & Slamovits, C. H. The Origin of Mitochondrial Cristae from Alphaproteobacteria. *Mol. Biol. Evol.* **34**, 943–956 (2017).
50. Gutiérrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nature Ecology & Evolution* **2**, 1700–1708 (2018).
51. Saghaï, A. *et al.* Comparative metagenomics unveils functions and genome features of microbialite-associated communities along a depth gradient. *Environ Microbiol* **18**, 4990–5004 (2016).
52. Saghaï, A. *et al.* Metagenome-based diversity analyses suggest a significant contribution of non-cyanobacterial lineages to carbonate precipitation in modern microbialites. *Front Microbiol* **6**, 797 (2015).

53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
54. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
56. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).
60. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
61. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
62. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
63. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).
64. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997).
65. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
66. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
67. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
68. Kannan, S., Rogozin, I. B. & Koonin, E. V. MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. *BMC Evolutionary Biology* **14**, 237 (2014).
69. 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).
70. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
71. Ali, R. H., Bogusz, M. & Whelan, S. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. *Mol Biol Evol* **36**, 2340–2351 (2019).
72. de Vienne, D. M., Ollier, S. & Aguileta, G. Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol. Biol. Evol.* **29**, 1587–1598 (2012).
73. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).
74. Vaidya, G., Lohman, D. J. & Meier, R. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics* **27**, 171–180 (2011).
75. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol* **67**, 216–235 (2018).
76. Schrempf, D., Lartillot, N. & Szöllősi, G. Scalable Empirical Mixture Models That Account for Across-Site Compositional Heterogeneity. *Molecular Biology and Evolution* **37**, 3616–3631 (2020).

77. Lartillot, N. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* **21**, 1095–1109 (2004).
78. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
79. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).
80. Susko, E. Tests for Two Trees Using Likelihood Methods. *Molecular Biology and Evolution* **31**, 1029–1039 (2014).
81. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Molecular Biology and Evolution* **16**, 1114–1114 (1999).
82. Markowski, E. A comparison of methods for constructing confidence sets of phylogenetic trees using maximum likelihood. (Dalhousie University, 2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- SupplementaryMaterial20210523.docx