nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01638-2

Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria

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

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 are sister to the Alphaproteobacteria to the exclusion of MarineProteo1 and Magnetococcia. We also show that the ancestral presence of the crista-developing mitochondrial contact site and cristae organizing system (a mitofilin-domain-containing Mic60 protein) in mitochondria and the Alphaproteobacteria only supports their close relationship.

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 propose that mitochondria provided an excess of energy claimed to be required for the origin of eukaryotic complexity[4], whereas others suggest that mitochondrial symbiosis merely 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, in contrast, have for decades been known to be phylogenetically associated with the Alphaproteobacteria[1,9,10]. However, the precise relationship between mitochondria and the Alphaproteobacteria, or any of its subgroups, has been elusive and remains a matter of intense debate (for example, see refs. [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 (that is, 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[12,26]. 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 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 (for example, through multiple amino acid replacements) that was originally present in the few genes that mitochondria and alphaproteobacteria still share. Second, the Alphaproteobacteria are under-sampled and most of their 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 and 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 and C nucleotides and G, A, R and P amino acids)[29]. This

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

sort of compositional heterogeneity is often the cause of artefactual attractions among lineages with similar compositional biases in phylogenetic inference[31].

In this Article, 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 (compared with <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

Until now, 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[11,12,16–18,32]. These markers are not only few (for example, 24 genes and 6,649 sites in ref. [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 and Extended Data Fig. 1). Our expanded dataset comprises most marker proteins previously identified[11,19,20] and adds 56 new ones (Extended Data Fig. 1). Functional annotations show that these proteins have diverse functions within mitochondria (Fig. 1b and Supplementary Table 1). Most are involved in energy metabolism (for example, respiratory chain complex subunits) and protein synthesis (for example, ribosomal subunits) (Fig. 1b and Supplementary Table 1). 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 compared with those that are mitochondrion encoded and some that are both nucleus and mitochondrion encoded (Fig. 1a). Similarly, nucleus-encoded proteins also have a broader range of GARP/FIMNKY 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 FIMNKY amino acids (and their genes towards A + T). The expanded set of nucleus-encoded genes are expected to increase the 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 (for example, refs. [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 appears to be 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 Supplementary Table 2 for details). In addition, we also screened MAG collections released previously[11,33–39], as well as the Genome Taxonomy Database (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,d and Supplementary Table 3), including those recently reported[11,40] (Fig. 1d and Supplementary Table 3). The most novel MAGs include new members of the 'early diverging' MarineProteo1 clade whose genomes are estimated to be relatively small or moderate in size (1.37–2.90 Mbp) and variable in nucleotide composition but not heavily biased towards A + T (31.3–59.7% G + C) (Extended Data Fig. 2 and Supplementary Table 4). In addition, several novel MAGs for 'basal' members of the Rickettsiales were estimated to be larger in size (1.82–2.81 Mbp) and enriched in G + C (37.2–61.3% or ~49.7% G + C on average) relative to previously sampled members of this group (0.89–2.37 Mbp and 21.6–50% G + C or ~33.7% G + C on average in the Rickettsiaceae, Anaplasmataceae, Midichloriaceae and Arcanobacteraceae) (Extended Data Fig. 2 and Supplementary Table 4). The new alphaproteobacterial MAGs have moderate to high quality (according to criteria by refs. [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 and Supplementary Table 3); this suggests that the methods used here to recover MAGs were not biased towards those with certain features (for example, 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 suggests that these MAGs come from photosynthesizers in the diverse environments sampled (for example, microbial mats; Fig. 1d and Supplementary Table 3).

To address recent controversies[11,12,26], we assembled a dataset that includes a new set of 64 nucleus-encoded and 44 mitochondrion-encoded proteins (108 proteins in total and 33,704 amino acid sites; see above). Our dataset also comprises a wide taxon sampling with 12 mitochondria from diverse eukaryotes (from most 'supergroups'), and a broad set of 104 alphaproteobacteria that covers all major known lineages and maximizes phylogenetic diversity (subsampled from a set of more than 3,300 genomes to decrease computational burden; Methods). Importantly, our dataset incorporated several Rickettsiales species that have short branches and are less compositionally biased (Fig. 1d, Extended Data Fig. 2 and Supplementary Table 4), as well as novel representatives of the MarineProteo1 clade (Figs. 1d and 2a and Supplementary Table 4). Instead of relying on Magnetococcia, and Betaproteobacteria and Gammaproteobacteria as outgroups (as in refs. [11,12]), we only used the much closer Magnetococcia which has been consistently found to be sister to all other alphaproteobacteria (for example, refs. [11,12,20]). This was done to decrease potential artefactual attractions between the long mitochondrial branch and distant outgroups, a concern raised before[11,12,26]. 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[41]. 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 analysed 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 (for example, C10-60)[42]. Analyses on the untreated dataset (that is, without compositionally

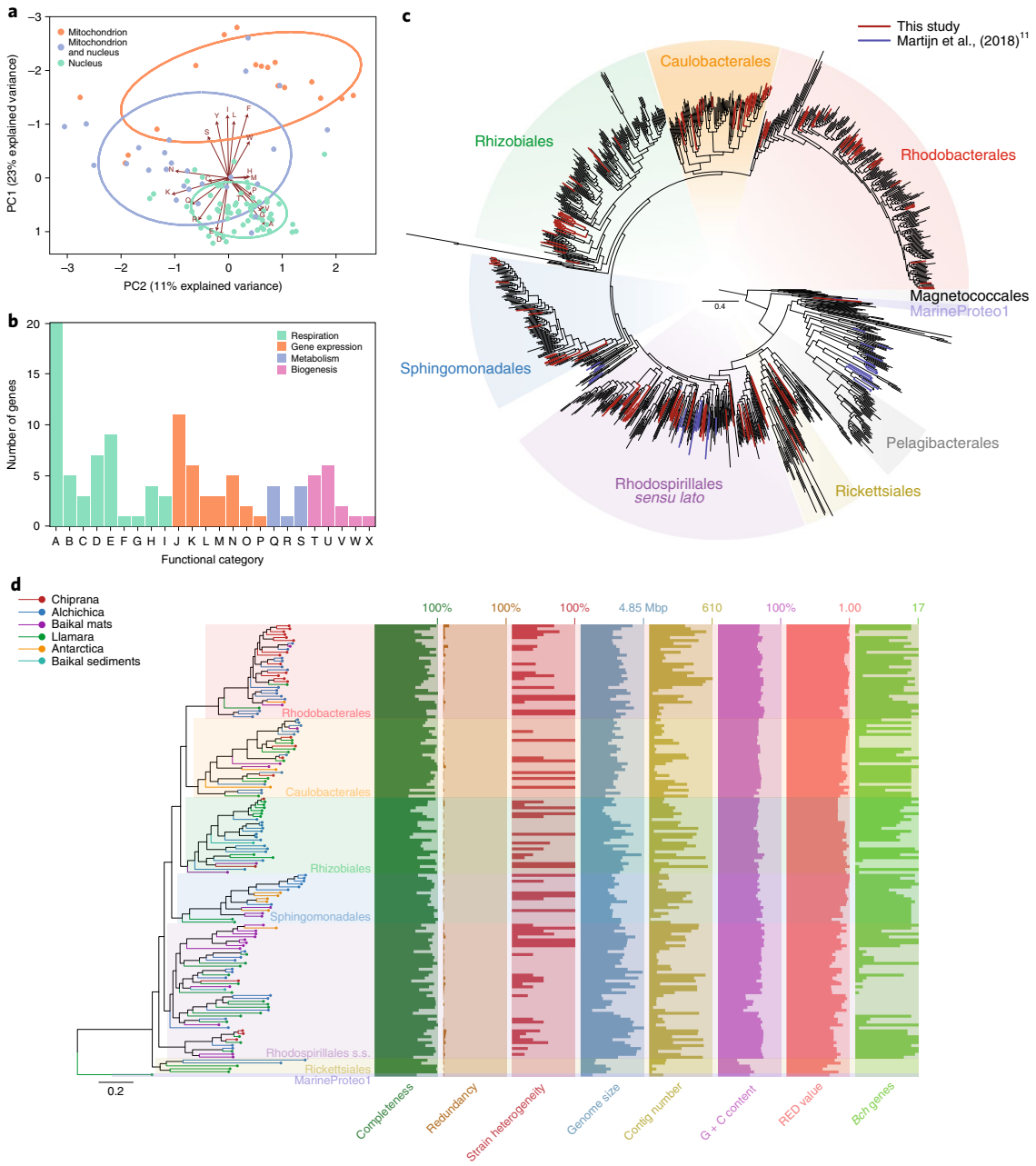


Fig. 1 | An expanded protein set and novel alphaproteobacterial MAGs from diverse environments. a, Principal component analysis (PCA) of amino acid compositions for each one of the 108 mitochondrial proteins of alphaproteobacterial origin used in this study. Light red, mitochondrion-encoded proteins; light blue, mitochondrion- and nucleus-encoded proteins; light green, nucleus-encoded proteins; 95% confidence ellipses follow the same colour code as proteins. This PCA analysis was inferred from protein alignments that contain only eukaryotes. **b**, Functional classification of the marker proteins of alphaproteobacterial origin in eukaryotes used for multi-protein 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 Supplementary Table 1). **c**, Phylogenetic tree of 154 novel MAGs reported here, the 45 MAGs reported by Martijn et al.[11], and 1,188 of maximally diverse alphaproteobacterial genomes in GTDB r89 database. Scale bar corresponds to 0.4 amino acid replacements per site. **d**, Phylogenetic tree for the 154 alphaproteobacterial MAGs reconstructed from diverse metagenomes sequenced in this study and summary of major features for each MAG. Scale bar corresponds to 0.2 amino acid replacements per site. Relative evolutionary divergence (RED) value.

heterogeneous sites removed) placed mitochondria as sister to all of the Alphaproteobacteria with maximum support, that is, 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

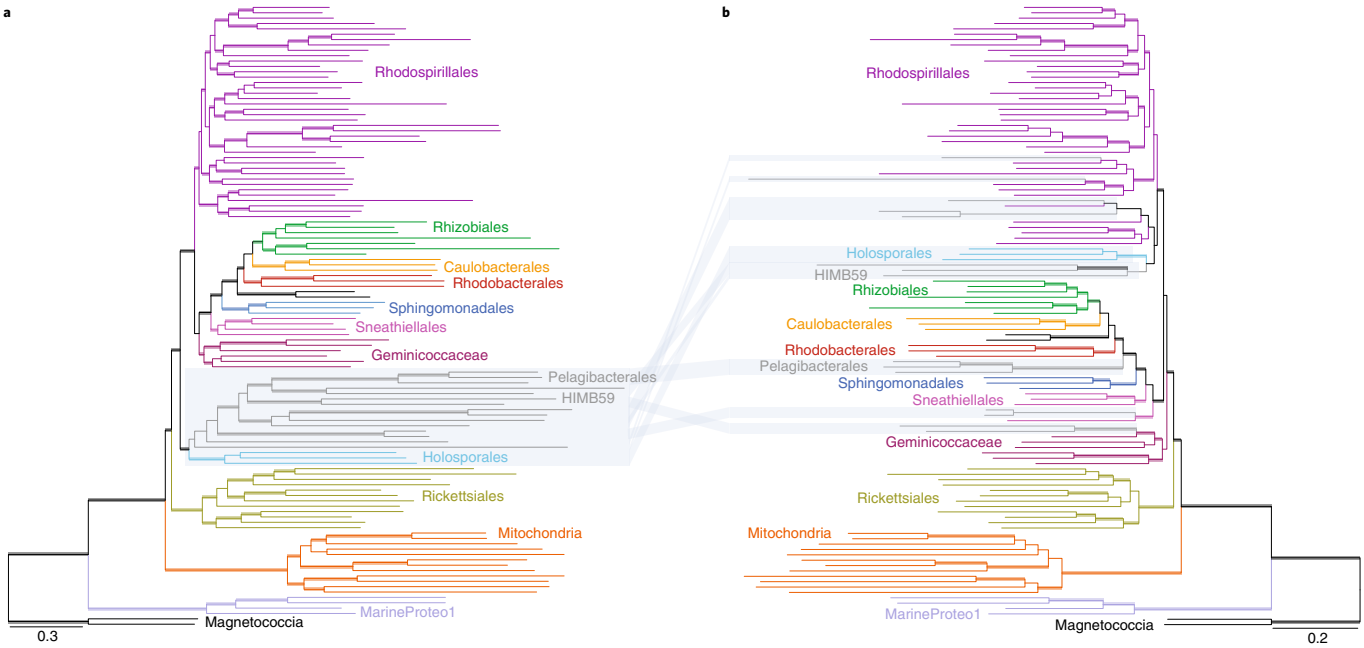


Fig. 2 | Phylogenetic trees of the Alphaproteobacteria and mitochondria from site-heterogenous analyses of untreated and compositionally homogenized datasets through site removal. a, Phylogenetic tree for the Alphaproteobacteria and mitochondria derived from a site-heterogeneous analyses of an untreated 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. Removal of this amount of sites minimizes the variation of GARP/FIMNKY amino acid ratios across taxa (Supplementary Table 5). The taxonomic labels follow the higher-level taxonomy outlined in ref. [29]. Thickened branches represent branch support values of >90% Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) and >90% UltraFast Bootstrap 2 with NNI optimization (UFBoot2 + NNI).

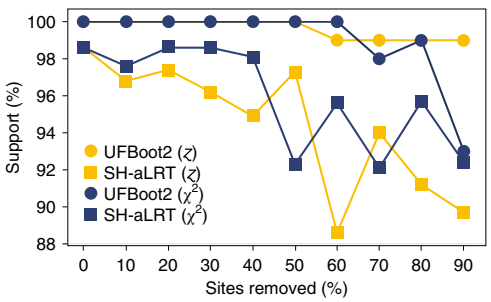


Fig. 3 | Branch support variation for the placement of mitochondria outside of the Alphaproteobacteria throughout the progressive removal of compositionally heterogenous sites. Branch support values are SH-aLRT and UFBoot2 + NNI, and compositionally heterogeneous sites were removed according to the ζ and χ^2 metrics. Support for the branch that groups mitochondria with all alphaproteobacteria (but excludes MarineProteo1 and the Magnetococcia) is always maximum (that is, 100% SH-aLRT /100% UFBoot2 + NNI; Mendeley Data[43]).

long-branching species (Fig. 3, Mendeley Data[43]) 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 (Figs. 2b and 3).

As nucleus-encoded and mitochondrion-encoded proteins display different amino acid compositional patterns (Fig. 1a), we also analysed these two protein sets separately. Proteins that are both mitochondrion and nucleus encoded were included in a mitochondrion-encoded protein dataset (M1) as most of these are encoded in mitochondrial genomes (Supplementary Table 1). An additional mitochondrion-encoded protein dataset (M2) was created by replacing the nucleus-encoded protein sequences with missing data. Whereas nucleus-encoded proteins unambiguously supported the Alphaproteobacteria-sister hypothesis across all analyses (Extended Data Fig. 3a, Mendeley Data[43]), both mitochondrion-encoded protein datasets showed decreased support for this hypothesis as compositionally heterogeneous sites were removed (Extended Data Fig. 3b,c, Mendeley Data[43]). However, neither of the two alternative topologies favoured by mitochondrion-encoded proteins (that is, Alphaproteobacteria-sister, or Caulobacteridae-sister where mitochondria are sister to all alphaproteobacteria except the Rickettsiales) was consistently and strongly supported (Extended Data Fig. 3b,c, Mendeley Data[43]). This suggests that mitochondrion-encoded proteins may have a more equivocal or less phylogenetic signal. We hypothesize that this could be the consequence of extreme compositional heterogeneity for primarily mitochondrion-encoded proteins in our dataset, and mutational saturation in mitochondrial genomes. Unlike in many previous studies[11,12,19,20], we did not find support for the Rickettsiales-sister hypothesis in any of our analyses (Mendeley Data[43]). We believe that the inclusion of new species of the Rickettsiales with less compositionally biased genomes might have decreased support for the Rickettsiales-sister topology. Indeed, replacing the Rickettsiales species in our dataset for a set of derived and compositionally biased Rickettsiales used by previous studies[11,12] recovered the Rickettsiales-sister hypothesis before >30% of the most compositionally heterogeneous sites were removed (Extended Data Fig. 4).

Until now, all studies have relied exclusively on either site-homogenous or purely site-heterogeneous models (for example,

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 GARP/FIMNKY amino acid ratio that is driven by variation in G + C versus A + T nucleotide content[29]. Our new branch-heterogeneous model, GFmix, models the variation in the ratio of GARP/FIMNKY amino acid frequencies across the phylogenetic tree in combination with conventional site-heterogeneous models (for example, C10-60, MAM and UDM models). Briefly, this model requires a rooted tree, and introduces a new parameter that represents the GARP/FIMNKY ratio for every branch in a tree that is based on the amino acid compositions of all taxa that descend from that branch (see 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, $Q^{(ce)}$, for each mixture class c for the given branch e. 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,b and Extended Data Fig. 5). The first tree set was inferred from the untreated dataset (108 proteins, 33,704 sites), whereas the second tree set was inferred from a compositionally homogenized dataset through site removal (108 proteins, 16,029 sites); the latter dataset minimized the differences of GARP/FIMNKY amino acid ratios among taxa (Supplementary Table 5). These two tree sets might thus correspond to two distinct regions in 'tree space' where compositional attractions abound or have been decreased, respectively (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,b and Extended Data Fig. 5). Furthermore, we also grouped proteins into partitions according to distances calculated based on their GARP/FIMNKY compositional disparity (Extended Data Fig. 6). Our analyses show that likelihoods estimated under the MAM60 + GFmix model improved significantly when compared with conventional site-heterogeneous models (Fig. 4 and Supplementary Table 6; likelihood ratio test (LRT) $P = 0$); model fit was improved even more when the proteins were grouped into ten separate partitions according to GARP/FIMNKY compositional disparity (Fig. 4 and Supplementary Table 6; LRT $P = 0$). 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 (for example, Pelagibacterales, Holosporaceae) is disrupted (that is, those trees recovered from compositionally homogenized datasets through site removal based on the ζ metric; Fig. 4 and Supplementary Table 6). This suggests that the removal of sites with extreme ζ scores effectively decreases overall compositional heterogeneity and potential artefacts.

The top three trees often favoured by the MAM60 + GFmix model (that is, those with the highest likelihoods) have mitochondria in adjacent branches: Alphaproteobacteria-sister (trees A11 and B9 in Fig. 2a,b), Rickettsiales-sister (trees A5 and B4 in Fig. 2a,b) and Caulobacteridae-sister (trees A10 and B8 in Fig. 2a,b)[29,47]. Bonferroni-corrected χ^2 topology tests show that the optimal trees that display the Alphaproteobacteria-sister relationship are significantly better than all other trees with alternative positions for mitochondria in almost all analyses (Fig. 4 and Supplementary Tables 6–9). The mitochondrion-encoded proteins, however, showed a more equivocal signal: the Caulobacteridae-sister and Alphaproteobacteria-sister topologies were not significantly different in MAM60-GFmix analysis of either mitochondrion-encoded protein dataset (Supplementary Tables 8 and 9). For example, the Caulobacteridae-sister relationship was slightly favoured in the partitioned MAM60-GFmix analysis of the M2 dataset, but the Alphaproteobacteria-sister topology was not rejected by Bonferroni-corrected χ^2 topology tests ($P > 0.01$; Supplementary Table 9). 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 ($P < 0.005$; Supplementary Tables 6–9). Overall, most of our distinct phylogenetic approaches show support for the Alphaproteobacteria-sister hypothesis.

Discussion

We have found substantial support for the Alphaproteobacteria-sister hypothesis that has the mitochondrial lineage as the closest sister to all currently sampled alphaproteobacteria[11]. 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.[12] was prone to certain artefacts. In an effort to choose less compositionally biased (that is, 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 Supplementary Figs. 31–48). Similarly, *Anaplasma*, *Neorickettsia* and *Wolbachia* (Anaplasmataceae) are among the longest branches in the Rickettsiales (see their Supplementary Fig. 50; see also our Extended Data Fig. 2). All these species are probably secondarily, and not ancestrally, less compositionally biased, that is, they evolved from species with A + T-rich genomes[48] (Extended Data Fig. 2). Moreover, their analyses were based on a rather small dataset that comprised only 18 or 24 mitochondrion-encoded proteins (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 (for example, see their Figs. S31–40), and an artefactual attraction between mitochondria, the Rickettsiales and the Fast-Evolving MAG (FEMAG) I and II groups (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 composed of both the Rickettsiales and the Holosporaceae[20]. These hypotheses implied that the mitochondrial ancestor may have been an intracellular parasite. In this scenario, the ancestor of mitochondria changed its function from an energy parasite to an ATP-producing respiratory organelle throughout its early evolution[18–21]. The finding that mitochondria are no longer phylogenetically associated with 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 or moderate in size, but not necessarily compositionally biased, and suggest that these alphaproteobacteria might be reduced or physiologically specialized (Extended Data Fig. 2 and Supplementary Table 4).

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

a

MAM60 (1 partition)		MAM60 + GFmix (1 partition)		MAM60 (10 partitions)		MAM60 + GFmix (10 partitions)	
Tree	LR	Tree	LR	Tree	LR	Tree	LR
A11	* 0	B9	* 0	A11	* 0	B9	* 0
A10	50.8	B8	54.6	A10	88.3	B8	91.6
A5	65.7	B4	63.6	A5	100.0	B4	101.1
A8	125.3	B3	227.9	A8	174.9	B3	319.6
A9	177.4	B7	240.1	A9	238.8	B7	321.6
A4	253.7	A11	261.6	A3	299.8	B1	392.0
A3	263.4	B1	282.6	A4	302.2	B2	431.8
A1	318.4	A10	307.4	A1	413.1	A11	1,117.4
A13	332.4	A5	321.4	A13	428.9	A10	1,209.7
A14	351.8	B2	331.1	A14	440.7	A5	1,221.0
A2	363.7	A8	429.0	A2	466.1	A8	1,369.2
B9	1,251.6	A9	458.4	B9	587.5	A9	1,393.5
B8	1,297.5	A4	484.5	B8	662.1	A4	1,410.5
B4	1,306.4	A3	517.5	B4	677.2	A3	1,422.3
B7	1,486.2	A1	585.9	B3	887.9	A1	1,558.5
B3	1,486.5	A13	605.0	B7	892.8	A13	1,559.8
B1	1,535.8	A14	612.2	B1	954.7	A14	1,570.3
B2	1,583.0	A2	630.7	B2	990.3	A2	1,622.7
A6	1,682.1	B5	1,656.1	A6	1,659.4	B5	1,660.9
A7	1,685.0	B6	1,659.4	A7	1,668.1	B6	1,671.6
A12	1,685.0	B10	1,659.4	A12	1,668.2	B10	1,671.7
B5	2,937.5	A6	1,908.0	B5	2,264.3	A6	2,762.5
B6	2,940.7	A7	1,911.2	B6	2,272.9	A12	2,764.6
B10	2,940.7	A12	1,911.2	B10	2,272.9	A7	2,766.5

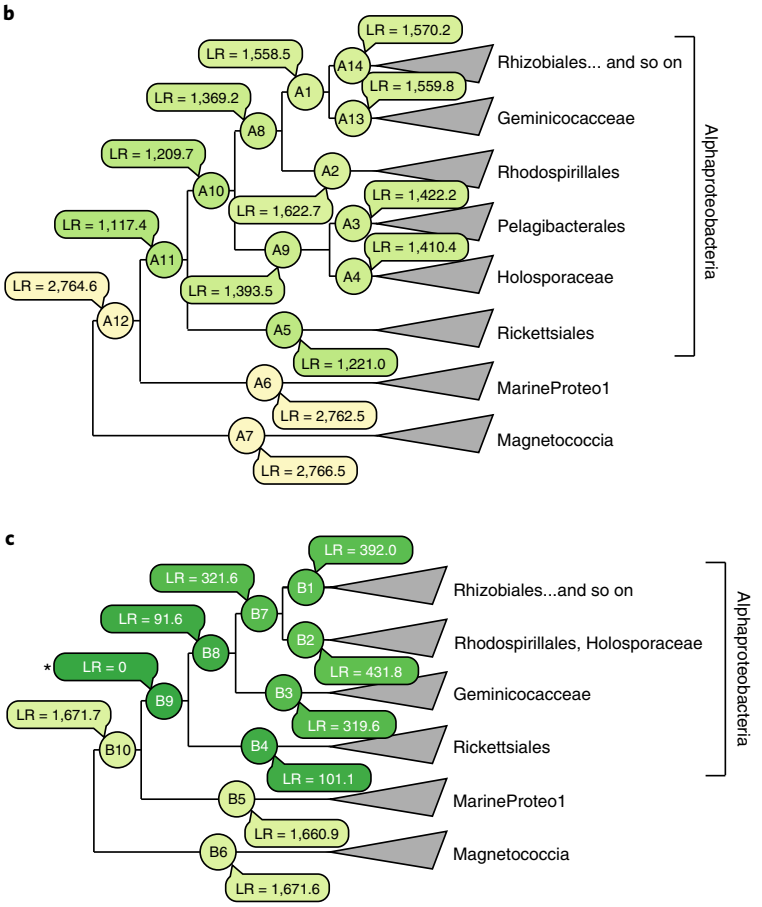


Fig. 4 | Support by the site-and-branch-heterogeneous MAM60 + GFmix model for several alterative placements of mitochondria relative to the Alphaproteobacteria. a, Likelihood ratio (LR) statistic for two sets of trees for several alternative placements of mitochondria (A1–A14 and B1–B10) under conventional site-heterogeneous (MAM60) and site-and-branch-heterogeneous (MAM60 + GFmix) models and two partition schemes (either one or ten partitions; see Methods for details). LR values are ordered decreasingly and coloured sequentially from green to yellow. For all four models, all trees other than the maximum-likelihood one were rejected with $P < 0.0001$ according to Bonferroni-corrected χ^2 tests (white asterisks). **b**, Alternative positions for mitochondria and associated LR values in the tree derived from a site-heterogenous analysis of the untreated dataset (trees A1–A12). **c**, Alternative positions for mitochondria and associated LR values in the tree derived from a site-heterogeneous analysis of a dataset from which 50% of its most compositionally heterogeneous sites were removed according to the ζ metric (trees B1–B10). See Supplementary Tables 6–9 and Extended Data Fig. 5 for all tree topologies and datasets tested.

have moved towards overcoming this major obstacle. Our newly developed and implemented site-and-branch-heterogenous model enabled us 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[11] is robust and supported by the bulk of the data. However, we caution that the phylogenetic signal preserved in mitochondrion-encoded proteins might be weak and ambiguous. The recovery of the Rickettsiales-sister relationship in previous studies[12] was most likely the result of little phylogenetic signal and long-branch attraction due to the presence of Rickettsiales species with fast-evolving and compositionally biased genomes, as we showed here. Therefore, we suggest that it is currently best to view mitochondria as an early offshoot of the alphaproteobacterial lineage that diverged just before the diversification of known extant groups. The closer phylogenetic affiliation of mitochondria to the Alphaproteobacteria, rather than to any other proteobacterial group, is suggested by the short internal branch lengths between mitochondria and Alphaproteobacteria (Fig. 2a,b), and supported by the shared presence of the mitochondrial contact site and cristae organizing system (that is, a mitofilin-domain-containing Mic60) in mitochondria and the Alphaproteobacteria. A large-scale phylogenetic profiling of the mitofilin-domain-containing Mic60 in hundreds of species of the Proteobacteria, expanding on previous analyses[49,50], reveals that this protein is unique to the Alphaproteobacteria and is thus absent in the MarineProteo1 clade, Magnetococcia, and Gammaproteobacteria and Zetaproteobacteria (Extended Data Figs. 2 and 7 and Supplementary Table 4). Future efforts should focus on exploring diverse environments for unknown and extant alphaproteobacterial lineages that may be more closely related to mitochondria.

Methods

Metagenome sequencing and MAG assembly. Samples collected from (1) microbial mats in the Salada de Chiprana (Spain, December 2013), Salar de Llamara[51], Lakes Bezymyannoe and Reid (Antarctica, January 2017) and several hot springs around Lake Baikal (Southern Siberia, July 2017), (2) microbialites in Lake Alchichica[52] and (3) sediments in Lake Baikal, were fixed in ethanol (>70%) in situ and stored at −20 °C as previously described[51]. Total DNA was purified from samples using the DNeasy PowerBiofilm Kit (QIAGEN) by following the manufacturer's guidelines.

DNA extracted from microbialite fragments was further cleaned using the DNeasy PowerClean Cleanup Kit (QIAGEN) as previously described[53]. DNA was quantified using a Qubit 3 fluorometer. DNA library preparation and sequencing were performed with an Illumina HiSeq2000 v3 (2 × 100 bp paired-end reads) by Beckman Coulter Genomics, and with an Illumina HiSeq2500 (2 × 125 bp paired-end reads) by Eurofins Genomics. A summary of the metagenomic libraries sequenced can be found in Supplementary Table 2.

Raw Illumina short reads from all sequenced Illumina paired-end libraries were quality-assessed with FastQC v0.11.7 and quality-filtered with Trimmomatic v0.36[54]. 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 (Supplementary Table 2), with metaSPAdes v3.10.0[55]. 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[56]. Contigs were binned using MaxBin v2.2.2, which relies on differential coverage across samples, tetranucleotide composition and single-copy marker genes[57]. The completeness and contamination of the bins reported by MaxBin v2.2.2 were assessed with CheckM v1.0.12[58]. Genome bins that were phylogenetically affiliated to the Alphaproteobacteria on the basis of 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[59]. In brief, reads were mapped to the final metaSPAdes co-assembly with Bowtie2 to obtain contig coverage values. DIAMOND searches[60] of predicted proteins against the NCBI GenBank nr database were done to assign taxonomic affiliations to each contig. CONCOCT2[61], 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, Southern Siberia, 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 v1.1.1[62]. 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 coverage values. Contigs were binned using MetaBAT v2.12.1[63], MaxBin v2.2.4[57] and CONCOCT2[61], and consensus bins were estimated with DAS Tool v1.1.0[64] (Supplementary Table 3).

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 v 2.7.1+[65] similarity searches of all proteins contained in the predicted proteomes of 13 representative eukaryotes were conducted against a database of 170 prokaryotes (136 bacteria and 34 archaea; Supplementary Table 10) with an *E*-value of 1×10^{-10}. Eukaryotic proteins (and the prokaryotic BLAST hits) were clustered into homologous families with a custom Perl script if more than 50% of their respective top 500 BLAST hits were identical. The corresponding datasets were subjected to several rounds of alignment, trimming, tree reconstruction and elimination of distant outgroups to refine the phylogenetic resolution. For this, they were aligned with the L-INS-I method of MAFFT v7.3.10[66], and then trimmed with BMGE v1.2[67] (-m BLOSUM30). Preliminary phylogenetic trees for each homologous protein family were inferred under the LG + G model in RAxML v8.2.12[68]. These trees were then sorted based on the criterion that eukaryotes form a monophyletic clade with alphaproteobacteria. Manual inspection of the trees then followed to remove paralogues and contaminants. For mitochondrion-encoded genes, mitochondrial clusters of orthologous genes (MitoCOGs)[69] 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[20] and Martijn et al.[11]. Our dataset encompassed most proteins from these other datasets, with few exceptions (Extended Data Fig. 1). The non-redundant and remaining candidate marker proteins comprising the union of these five datasets (Extended Data Fig. 1) were then further screened phylogenetically. Using a representative eukaryotic query (*Andalucia godoyi*) for each marker gene, BLASTp (-matrix BLOSUM45) searches were done against a database that comprises 107 diverse bacteria (representing 27 cultured phyla) and 23 diverse eukaryotes (representing 6 major groups) (Supplementary Table 11); eukaryotes were selected based on the availability of both mitochondrial and nuclear genomes or transcriptomes (Supplementary Table 12). Homologues were aligned with MAFFT v7.3.10 and the L-INS-I method, alignments trimmed with trimAl v1.4.rev15[70] (-automated1) and single-protein trees inferred with IQ-TREE v1.6.10[71] and the best model according to ModelFinder[72]. 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 v5.0 database and BLASTp searches. The final marker proteins set comprised 108 genes, of which 64 are exclusively nucleus encoded, 17 are exclusively mitochondrion encoded and 27 are both mitochondrion and nucleus encoded (Extended Data Fig. 1). The annotations confirm that all marker proteins are predicted to be localized to mitochondria in eukaryotes (Supplementary Table 1).

Dataset assembly. To increase taxon sampling as much as possible, MAGs reported by Anantharaman et al.[33], Graham et al.[34], Delmont et al.[35], Martijn et al.[11], Mehrshad et al.[36], Tully et al.[37], Tully et al.[38] and Parks et al.[39] were added to those reconstructed here (see Metagenome sequencing and MAG assembly; Supplementary Table 2). To improve the quality of our MAG selection, MAGs were analysed with the CheckM lineage workflow and those with quality values (completeness minus five times (5×) contamination) lower than 50 were discarded, as done previously by Parks et al.[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 into the GTDB-Tk software, using IQ-TREE v1.6.10[71] 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. To reduce computational burden, Treemmer v0.1b was then used to reduce the number of alphaproteobacterial taxa from the GTDB-TK tree while maximizing phylogenetic diversity[73]. 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. At this stage, a set of 161 taxa including 23 eukaryotes and 138 proteobacteria was kept for further phylogenetic screening (Supplementary Table 12).

To retrieve homologues from the above 161 taxa, PSI-BLAST v2.7.1+ searches (-matrix BLOSUM45; -evalue 1e04^{-4}) using representative mitochondrial (eukaryotic) query sequences for each marker protein were done against a database that comprised carefully selected predicted proteomes of alphaproteobacteria and mitochondria. PSI-BLAST searches were iterated until homologues could be retrieved for most taxa. Most proteins required only one or two iterations, except Atp4, which required a third PSI-BLAST iteration to retrieve a considerable number of homologues. To remove non-orthologous sequences, homologous protein sets were retrieved for each marker protein, aligned with MAFFT v7.3.10 L-INS-I and trimmed with trimAl v1.4.rev15 (-automated1), and trees were inferred with IQ-TREE v1.6.10 and the best-fitting model according to ModelFinder[72]. The individual 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 v7.3.10 and the L-INS-I method. To increase phylogenetic signal by removing poorly aligned and non-homologous aligned regions, Divvier v1.0 was used with the -partial and -mincol options[74]. Only sites with more than 10% of data were retained. To reduce incongruency among proteins due to, for example, lateral gene transfer, Phylo-MCOA v1.4[75] was employed on single-protein trees with UFBoot2+NNI as branch support which were inferred with IQ-TREE v1.6.10 and the best-fitting model as identified by ModelFinder[71,72]. Single-protein alignments were concatenated with SequenceMatrix v1.8[76]. To reduce further computational burden, the set of 161 taxa was manually reduced to 116 taxa containing a single outgroup (Magnetococcia) (Supplementary Table 13). The final dataset used for phylogenetic analyses comprised 116 taxa and 108 proteins (33,704 amino acid sites), of which 25.46% represented missing data (Supplementary Tables 14–16; see also figshare[77]).

Phylogenetic analyses using site-heterogeneous models. For multi-protein phylogenetic analyses on the supermatrix, trees were first inferred in IQ-TREE v1.6.10 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[78]. 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[42]. This site-heterogeneous mixture model is directly inferred from the dataset analysed and therefore is more specific than the general C10-60 mixture models. To account for more than 60 (for example, 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[79]. The software FunDi was used to estimate functionally divergent sites in the branch that separates the mitochondrial lineage from all other taxa[41]. Sites with a probability >0.5 of being functionally divergent were removed. Progressive removal of compositionally heterogeneous sites was performed according to the ζ and the χ^2 metrics/methods as described previously[11,29,44]. Both metrics are designed to estimate compositional heterogeneity per site on the basis of different criteria.

Bayesian analyses were conducted with PhyloBayes MPI v1.8 using the CAT-LG+G4 model[80,81]. 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[43]).

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

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

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 $\boldsymbol{\theta}$ is the vector of other adjustable parameters (branch lengths, α shape parameter and tree topology) in the model. To model shifts in the relative frequencies of the amino acids GARP (specified by G+C-rich codons) and FIMNKY (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 GARP to FIMNKY amino acids; that is, $b := \pi_G/\pi_F$ 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 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:

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

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

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

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

Partitioning of supermatrices for likelihood calculations under the GFmix model. The foregoing framework assumes that, for each aligned protein in a given concatenated dataset, the GARP/FIMNKY ratios (b_e) for every branch in the tree will be similar. However, for our data matrix, this assumption is not true as different proteins show different degrees of GARP/FIMNKY amino acid variation across taxa depending on the location of the corresponding protein (for example, nucleus encoded versus mitochondrion encoded) and degree of conservation. For this reason, we clustered the proteins in our dataset into groups in the following way. For each protein v and each taxon t, we calculated the GARP/FIMINKY ratio, $b_v^{(t)} = \pi_G^{(t)}/\pi_F^{(t)}$. Then, we calculated the overall distance between these ratios for every pair of proteins u and v in the data matrix as $d_{u,v} = \sum_t \left| b_v^{(t)} - b_u^{(t)} \right| / N_{u,v}$ where $N_{u,v}$ is the total number of taxa for which sequences were available for both proteins (this normalization accounts for the differing amounts of missing data for different proteins). The proteins were then clustered on the basis of $d_{u,v}$ distances using the UPGMA algorithm in MEGA-X[82], and clusters were chosen as a computationally tractable number of partitions for further analysis. Ten protein clusters (partitions) were chosen for the combined dataset, and five protein clusters (partitions) were chosen for each the nucleus-encoded and mitochondrion-encoded protein datasets (Extended Data Fig. 6). The GFmix model was then applied to these partitions allowing for separate b_e values and branch lengths for each partition. The overall log-likelihoods for topologies were obtained as the sum of log-likelihoods of that topology over all partitions.

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

Phylogenetic analyses using the site-and-branch-heterogeneous GFmix model. For estimating log-likelihoods, two sets of topologies were generated by varying the placement of the mitochondrial lineage in the maximum-likelihood tree that derived from site-heterogeneous analyses (LG+PMSF(MAM60)+F+G4) of the untreated dataset and a compositionally homogenized dataset obtained by removing sites with extreme ζ scores. Six sets of topologies were produced in such a way for the combined, nucleus-encoded and mitochondrion-encoded protein datasets (Extended Data Fig. 5 and Supplementary Tables 6–9). Likelihood estimations under the site-heterogeneous LG+MAM60+F+G4 model were done with IQ-TREE v2 on the fixed topologies and the two Magnetococcia species (GCF_002109495 and GCA_002753665) as outgroup. Likelihood calculations under the site-and-branch-heterogeneous LG+MAM60+F+G4+GFmix model were done with the GFmix v1.0 software (see below) on the fixed topologies and the two Magnetococcia species (GCF_002109495 and GCA_002753665) as outgroup. The above likelihood estimations were done on both non-partitioned and partitioned dataset according to protein GARP/FIMNKY ratios (see above and Extended Data Fig. 6).

Topology testing using the Bonferroni-corrected χ^2 test. The topology test is a variation of the chi-squared test presented in Susko[83] that corrects for selection bias. The chi-squared test is a test of two trees. The null hypothesis $H_0 : \tau = \tau_0$ is tested against the alternative hypothesis $H_A : \tau = \tau_A$, where τ is the true topology. As a test statistic, it uses the LRS, which is defined as twice the difference between the maximized log likelihood when the true topology is τ_A and the maximized log likelihood for τ_0. It gives a P value $P(\tau_A) = P[\chi_d^2 > \text{LRS}]$, the probability that a chi-squared random variable with d degrees of freedom is greater than the observed LRS. Here the degrees of freedom, d, are determined as the number of branches that have been collapsed (that is, 0 in length) in the consensus tree representing both τ_0 and τ_A.

In the absence of a particular τ_A of interest, to test whether $H_0 : \tau = \tau_0$ can be rejected, we consider the alternative $H_A : \tau = \hat{\tau}$, where $\hat{\tau}$ is the maximum likelihood 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[84]. The Bonferroni approach uses an 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-squared 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 ref. [83] for additional discussion). If the true tree is the consensus tree, then it is likely that the maximum likelihood 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; there is some possibility that a tree with fewer 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 adjust for multiple testing. As one final approximation, rather than using the usual Bonferroni-corrected P value, $n(A)\,P(\hat{\tau})$, 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 ref. [85].

Other analyses. Analyses done to obtain phylogenetic trees for display purposes were done as follows. Taxon subsampling to reduce computational burden and ease visualization was done with Treemmer v0.1b[73] (Fig. 1c and Extended Data Fig. 7). Datasets were assembled on the basis of bacterial (120 markers from GTDB-Tk for Fig. 1c,d), alphaproteobacterial (117 from GToTree v1.6.11[86] for Extended Data Fig. 2) or proteobacterial (119 markers from GToTree for Extended Data Fig. 7) single-copy marker genes. Removal of the 50% most compositionally heterogeneous sites based on their ζ scores was done as reported previously[29] (Fig. 1d). Phylogenetic analyses were done with IQ-TREE v1.6.10[71] and the LG4X model (-fast mode) (Fig. 1c,d, Supplementary Fig. 2 and Extended Data Fig. 7). Superposition of metadata layers and visualization was done in Anvi'o v7[59].

To search for bacteriochlorophyll enzymes, a set of 17 custom-made profile hidden Markov models (pHMMs) for the genes *bchB*, *bchC*, *bchD*, *bchE*, *bchF*, *bchG*, *bchH*, *bchI*, *bchJ*, *bchL*, *bchM*, *bchN*, *bchO*, *bchP*, *bchX*, *bchY* and *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 v3.3.2 suite using an *E*-value cut-off of 1×10^{-25}. To search for mitofilin-domain-containing *mic60* genes, the Pfam pHMM for mitofilin (PF09731) was used with its own GA (gathering) cut-off value.

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

Data availability

Sequencing data are deposited in NCBI GenBank under the BioProjects PRJNA315555, PRJNA438773, PRJNA754110, PRJNA754380, PRJNA752523 and PRJNA703749. Novel alphaproteobacterial MAGs and protein files (unaligned, aligned, and aligned and trimmed) are available at https://doi.org/10.6084/m9.figshare.14355845. Datasets and phylogenetic trees inferred in this study are available at https://doi.org/10.17632/dnbdzmjjkp.1.

Code availability

The GFmix model software is available at: https://www.mathstat.dal.ca/~tsusko/software.html

Received: 24 May 2021; Accepted: 29 November 2021;
Published online: 13 January 2022

References

1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **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. *Int. J. Biochem. Cell Biol.* **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. *Nat. Rev. Microbiol.* **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. *Proc. Natl Acad. Sci. USA* **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. *Nat. Ecol. Evol.* **4**, 1213–1219 (2020).
13. Viale, A. M. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett.* **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. *Sci. Rep.* **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? *Curr. Biol.* **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. *Nat. Commun.* **7**, 13219 (2016).
34. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* **12**, 1861–1866 (2018).
35. Delmont, T. O. et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **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**, 170203 (2018).
39. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
40. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
41. Gaston, D., Susko, E. & Roger, A. J. A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics* **27**, 2655–2663 (2011).
42. Susko, E., Lincker, L. & Roger, A. J. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. *Mol. Biol. Evol.* **35**, 1266–1283 (2018).
43. Muñoz-Gómez, S. A. et al. Additional Supplementary Data for 'Site-and-branch-heterogeneous analyses of an expanded dataset favor mitochondria as sister to known Alphaproteobacteria. *Mendeley Data* https://doi.org/10.17632/dnbdzmjjkp.1 (2021).
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. Smith, D. R. Updating our view of organelle genome nucleotide landscape. *Front. Genet.* **3**, 175 (2012).
49. 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. *Curr. Biol.* **25**, 1489–1495 (2015).
50. 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).
51. Gutiérrez-Preciado, A. et al. Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nat. Ecol. Evol.* **2**, 1700–1708 (2018).
52. 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).
53. 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).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
56. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
57. 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).
58. 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).
59. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
60. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
61. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
62. 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).
63. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
64. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
65. 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).
66. 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).
67. 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).
68. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
69. Kannan, S., Rogozin, I. B. & Koonin, E. V. MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. *BMC Evol. Biol.* **14**, 237 (2014).
70. 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).
71. 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).
72. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
73. Menardo, F. et al. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
74. Ali, R. H., Bogusz, M. & Whelan, S. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol. Biol. Evol.* **36**, 2340–2351 (2019).
75. 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).
76. 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).
77. Muñoz-Gómez, S. A. et al. Alignments for 108 mitochondrial proteins of alphaproteobacterial origin, and alphaproteobacterial MAGs from microbial mats, microbialites, and sediments. *figshare* https://doi.org/10.6084/m9.figshare.14355845.v2 (2021).
78. 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).
79. Schrempf, D., Lartillot, N. & Szöllősi, G. Scalable empirical mixture models that account for across-site compositional heterogeneity. *Mol. Biol. Evol.* **37**, 3616–3631 (2020).
80. Lartillot, N. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
81. 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).
82. 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).
83. Susko, E. Tests for two trees using likelihood methods. *Mol. Biol. Evol.* **31**, 1029–1039 (2014).
84. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114 (1999).
85. Markowski, E. *A Comparison of Methods for Constructing Confidence Sets of Phylogenetic Trees Using Maximum Likelihood*. MSc thesis, Dalhousie Univ. (2021).
86. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164 (2019).

Acknowledgements

S.A.M.-G. is supported by an EMBO Postdoctoral Fellowship (ALTF 21-2020). We thank B. Curtis (Dalhousie University) and D. Salas-Leiva (Dalhousie University) for assistance with scripts, W. Valencia (Harvard University) and C. Calderon (Rutgers University) for advice on Python and R, and A. Gutiérrez-Preciado (Université Paris-Saclay) for assistance with uploading data to NCBI GenBank. This work was supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation grants 735923LPI (https://doi.org/10.46714/735923LPI) awarded to A.J.R. and GBMF9739 (https://doi.org/10.37807/GBMF9739) awarded to P.L.G., and Discovery Grants from the Natural Sciences and Engineering Research Council of Canada awarded to A.J.R., E.S. and C.H.S.

Author contributions

S.A.M.-G.: conceptualization, methodology, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualization, project administration, funding acquisition. E.S.: methodology, software, writing—review and editing. K.W.: validation, data curation, writing—review and editing. L.E.: resources, writing—review and editing. C.H.S.: resources, supervision, writing—review and editing, funding acquisition. D.M.: resources, writing—review and editing, funding acquisition. P.L.-G.: resources, writing—review and editing, funding acquisition. A.J.R.: conceptualization, methodology, validation, resources, supervision, project administration, writing—review and editing, funding acquisition.

Competing interests

The authors declare no competing interests.

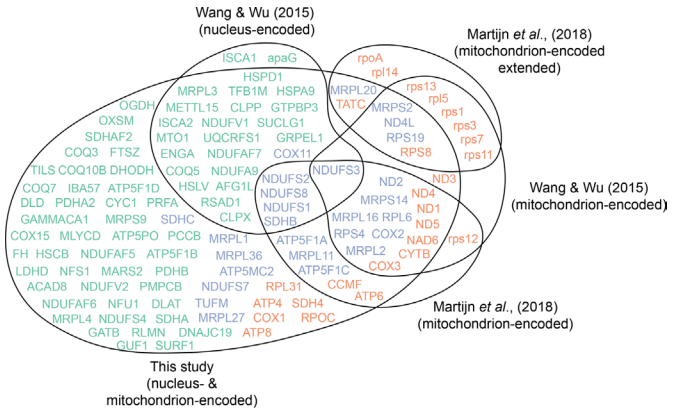
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01638-2.

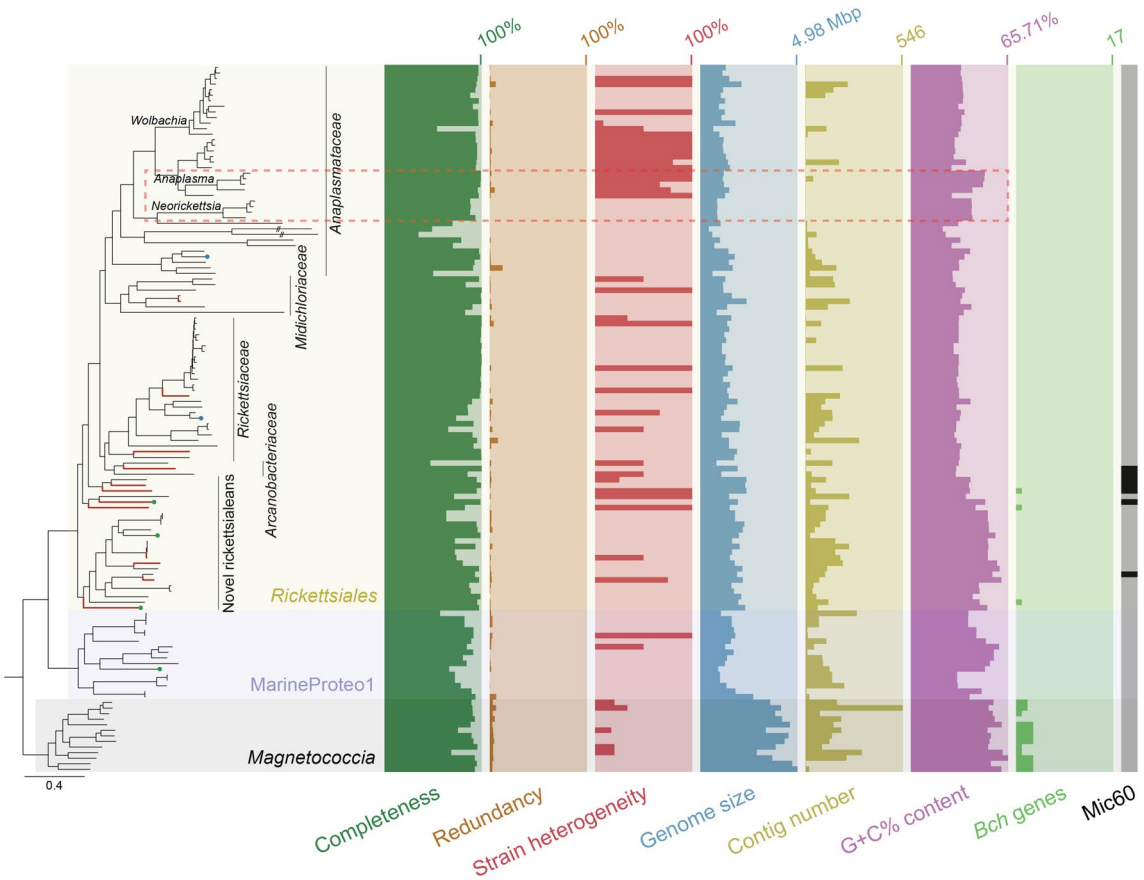
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01638-2.

Correspondence and requests for materials should be addressed to Sergio A. Muñoz-Gómez or Andrew J. Roger.

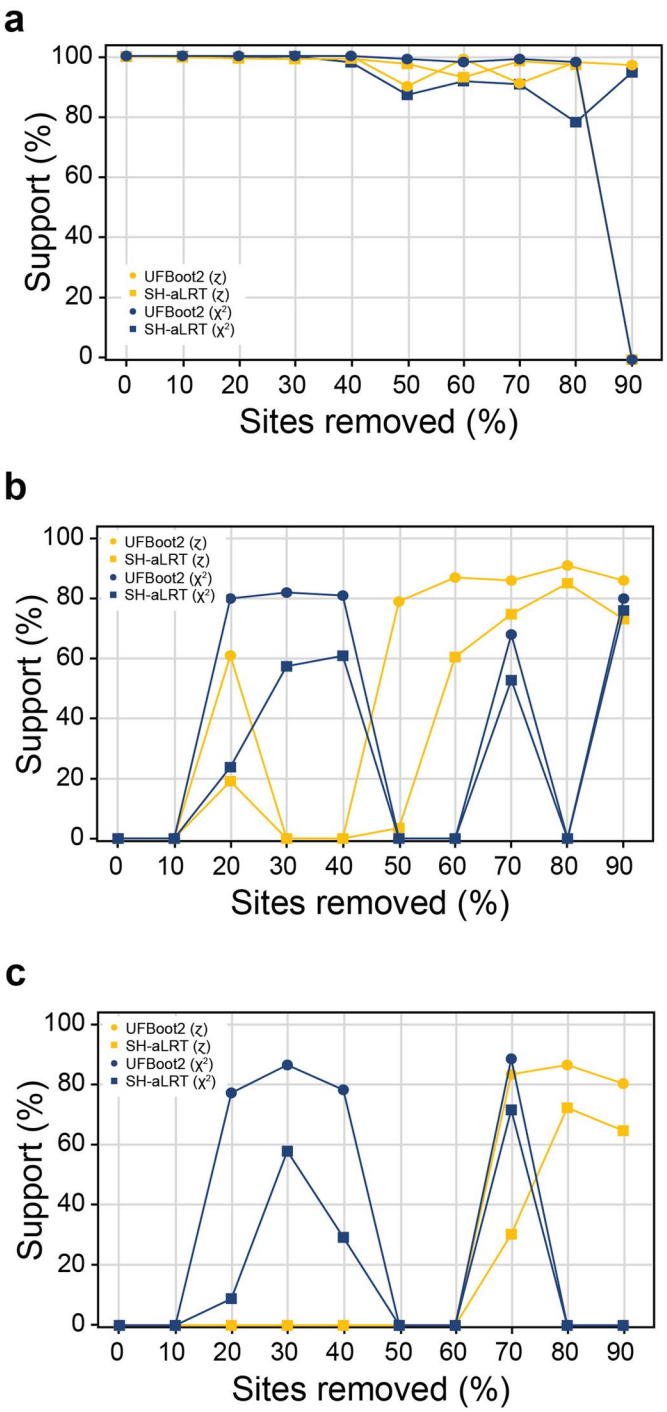
Peer review information *Nature Ecology & Evolution* thanks the anonymous reviewers for their contribution to the peer review of this work.



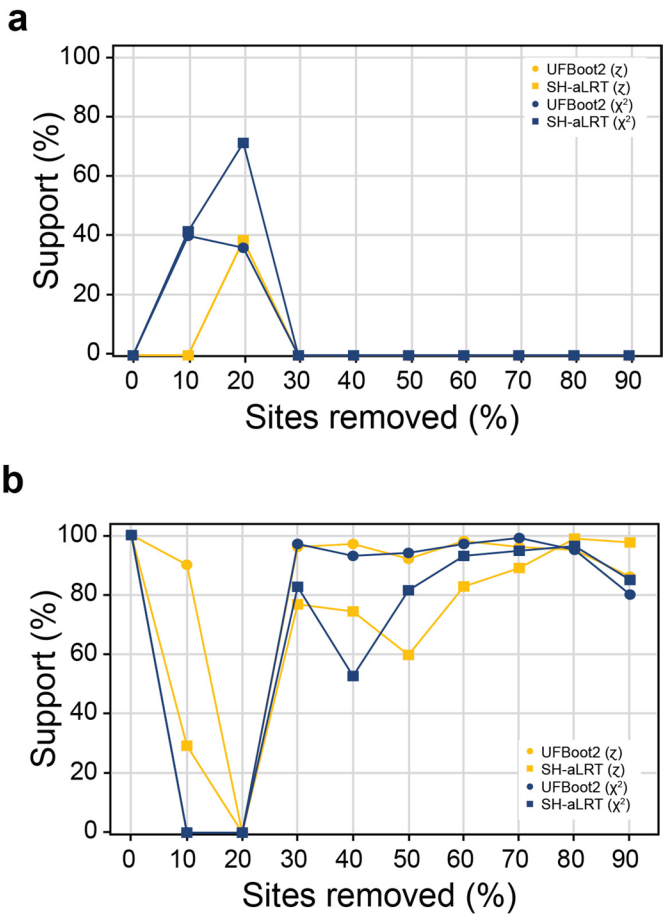
Extended Data Fig. 1 | Euler diagram that shows the relationships between recent phylogenomic sets of proteins used to address the phylogenetic placement of mitochondria. Datasets include those comprised of mitochondrion- and nucleus-encoded proteins in the studies Wang and Wu [20], Martijn et al. [11], and this study. Nucleus-encoded proteins are in green, mitochondrion-encoded proteins in red, and both nucleus- and mitochondrion-encoded proteins in blue. Gene/protein names mostly follow the human gene nomenclature.



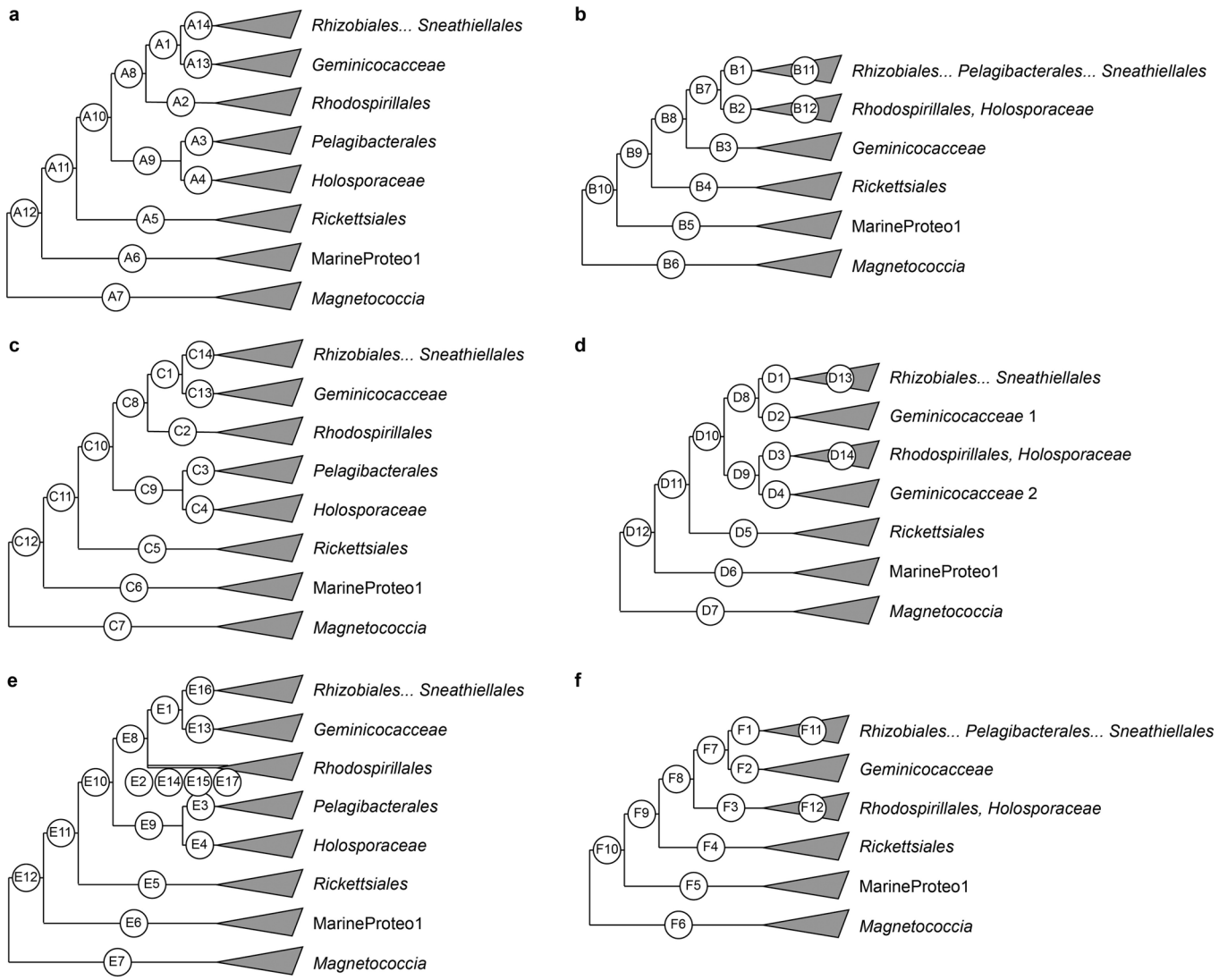
Extended Data Fig. 2 | Summary of features for novel MAGs that belong to the MarineProteo1 clade and the *Rickettsiales*. Branches highlighted in red show taxa used for phylogenetic analyses in this study. The dashed rectangle points to the secondary higher G + C% content of the genera *Anaplasma* and *Neorickettsia* in the family *Anaplasmataceae*. The *Magnetococcia* is at the base of the tree as an outgroup.



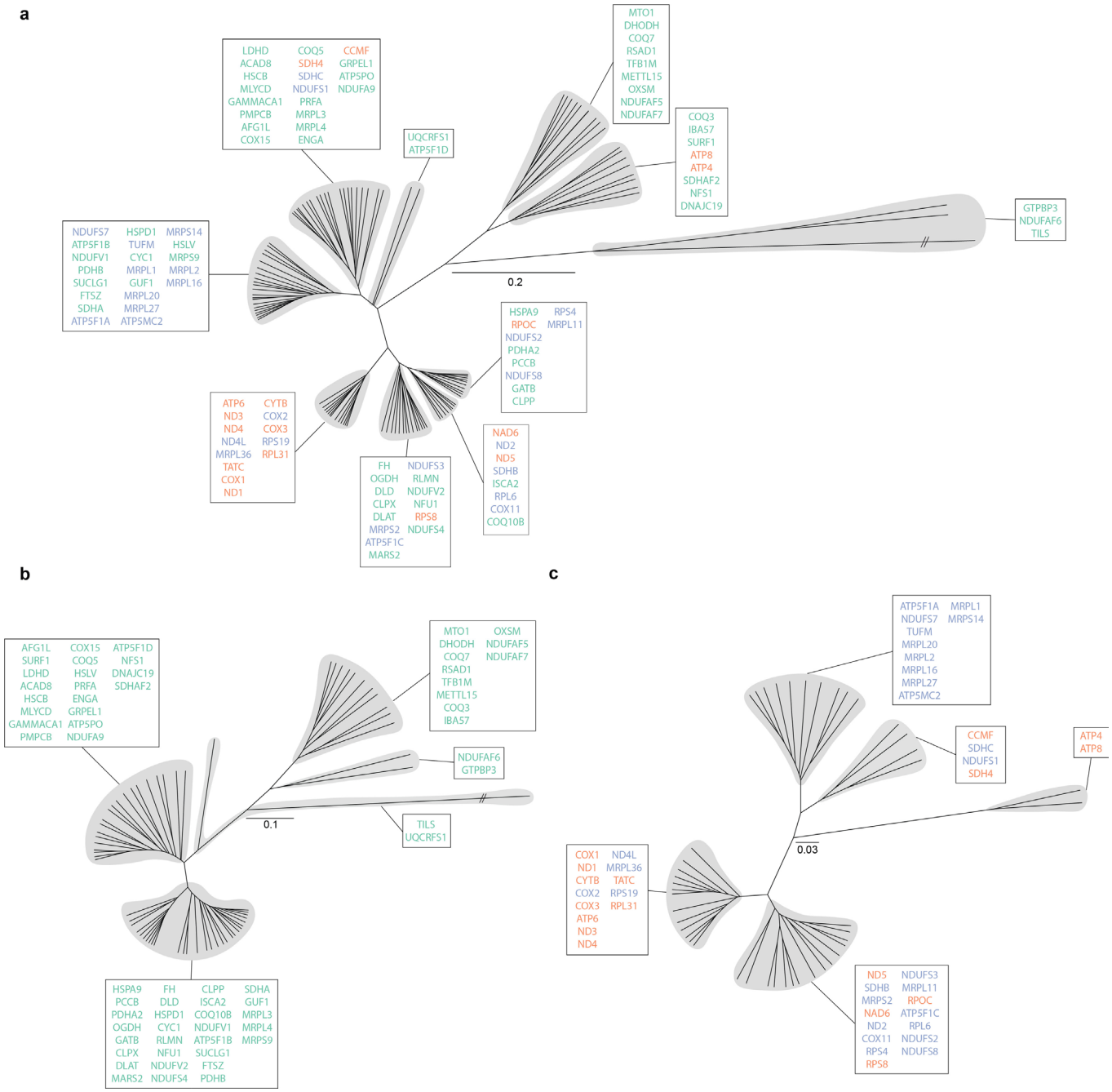
Extended Data Fig. 3 | Branch support variation for the placement of mitochondria outside of the *Alphaproteobacteria* throughout the progressive removal of compositionally heterogenous sites. Branch support values are SH-aLRT and UFBoot2+NNI and the removal of compositionally heterogeneous sites was done 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 maximal (i.e., 100% SH-aLRT /100% UFBoot2+NNI). (a) Nucleus-encoded protein dataset. (b) Mitochondrion-encoded protein M1 dataset. (c) Mitochondrion-encoded protein M2 dataset.



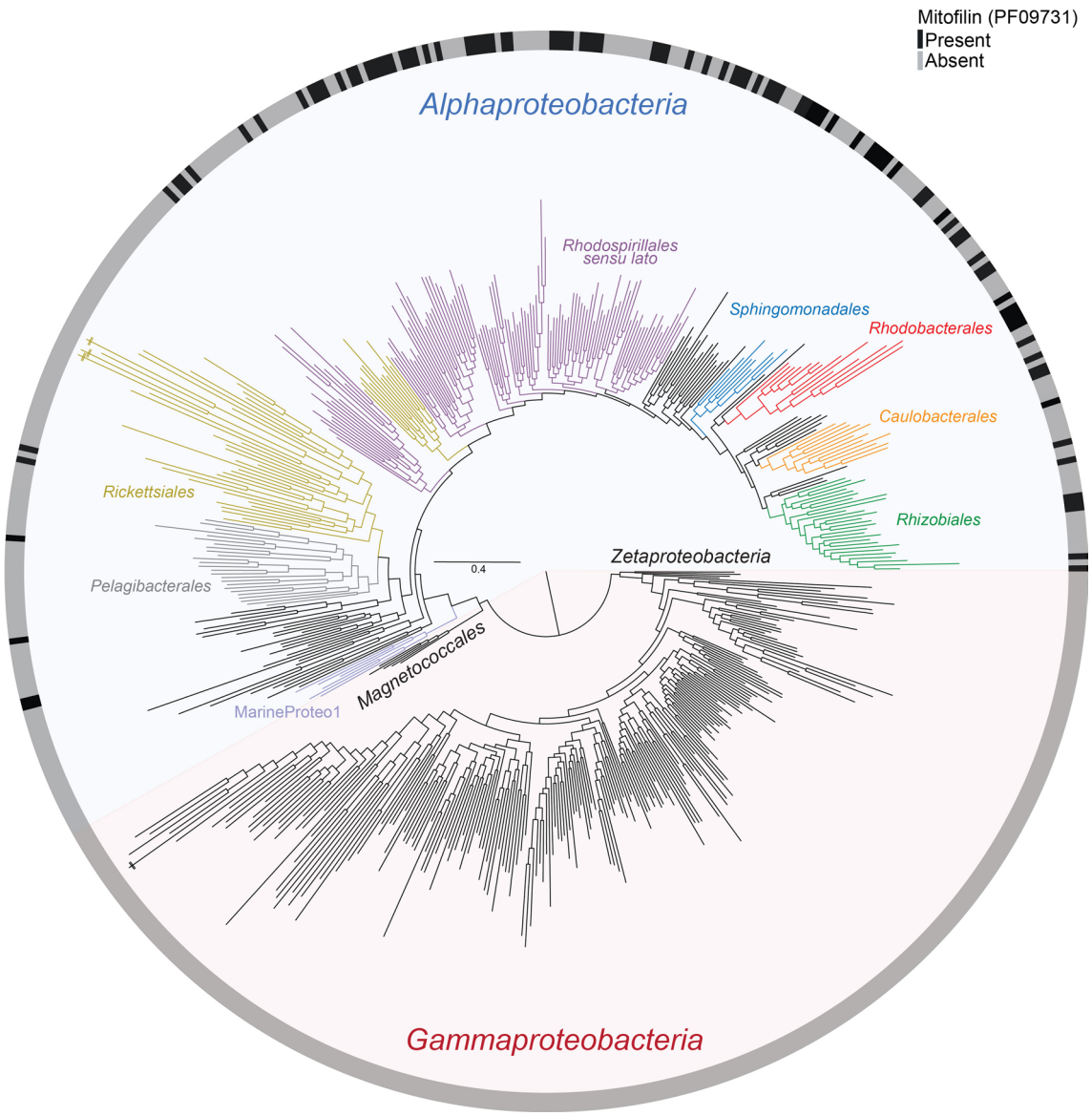
Extended Data Fig. 4 | Branch support variation for the placement of mitochondria when derived and compositionally biased *Rickettsiales* are included throughout the progressive removal of compositionally heterogenous sites. Branch support values are SH-aLRT and UFBoot2+NNI and the removal of compositionally heterogeneous sites was done according to the ζ and χ^2 metrics. (a) *Alphaproteobacteria*-sister topology. Support for the branch that groups mitochondria with all alphaproteobacteria (but excludes MarineProteo1 and the *Magnetococcia*) is always maximal (i.e., 100% SH-aLRT /100% UFBoot2+NNI). (b) *Rickettsiales*-sister topology.



Extended Data Fig. 5 | Schematic tree topologies used for calculating likelihood values using the MAM60 + GFmix model. (a) Tree topologies derived from analyses of the untreated dataset of mitochondrion-, and nucleus-encoded proteins. (b) Tree topologies derived from analyses of a compositionally homogenized dataset of mitochondrion-, and nucleus-encoded proteins. (c) Tree topologies derived from analyses of the untreated dataset of nucleus-encoded proteins. (d) Tree topologies derived from analyses of a compositionally homogenized dataset of nucleus-encoded proteins. (e) Tree topologies derived from analyses of the untreated dataset of mitochondrion-encoded proteins. (f) Tree topologies derived from analyses of a compositionally homogenized dataset of mitochondrion-encoded proteins. Datasets were compositionally homogenized by removing the 50% most compositionally heterogeneous sites according to the ζ metric.



Extended Data Fig. 6 | UPGMAs dendrograms for G A R P/F I M N K Y distances among the marker proteins of alphaproteobacterial origin in eukaryotes used in this study. (a) Mitochondrion- and nucleus-encoded proteins. (b) Nucleus-encoded proteins. (c). Mitochondrion-encoded proteins. Nucleus-encoded proteins are in green, mitochondrion-encoded proteins in red, and both nucleus- and mitochondrion-encoded proteins in blue. Gene/protein names mostly follow the human gene nomenclature.



Extended Data Fig. 7 | Phylogenetic distribution of the Mitofilin-domain containing Mic60 in the Proteobacteria. The Mitofilin-domain containing Mic60, as defined by the Pfam pHMM Mitofilin PF09731, is phylogenetically restructured to the *Alphaproteobacteria* to the exclusion of MarineProteo1 clade and the *Magnetococcia*. This protein is also conspicuously absent in the Gamma- and *Zetaproteobacteria*.

nature portfolio

Corresponding author(s): Sergio A. Munoz-Gomez, Andrew J. Roger

Last updated by author(s): Nov 23, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All software used in this study are publicly available, and described in detail in the Methods section.
Data analysis	The GFmix model software is available at: https://www.mathstat.dal.ca/~tsusko/software.html

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data were deposited in NCBI GenBank under the BioProjects PRJNA315555, PRJNA438773, PRJNA754110, PRJNA754380, PRJNA752523, and PRJNA703749. Novel alphaproteobacterial MAGs and gene files (unaligned, aligned, and aligned and trimmed) are available at: DOI: https://dx.doi.org/10.6084/m9.figshare.14355845. Datasets and phylogenetic trees inferred in this study are available at: DOI: https://dx.doi.org/10.17632/dnbdzmjjkp.1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	N/A
Data exclusions	No data were excluded from the analyses.
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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