nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-020-1239-x

Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria

Lu Fan[1,2,3,7]✉, Dingfeng Wu[4,7], Vadim Goremykin[5,7], Jing Xiao[4], Yanbing Xu[4], Sriram Garg[6], Chuanlun Zhang[1,3], William F. Martin[6]✉ and Ruixin Zhu[4]✉

Though it is well accepted that mitochondria originated from an alphaproteobacteria-like ancestor, the phylogenetic relationship of the mitochondrial endosymbiont to extant Alphaproteobacteria is yet unresolved. The focus of much debate is whether the affinity between mitochondria and fast-evolving alphaproteobacterial lineages reflects true homology or artefacts. Approaches such as site exclusion have been claimed to mitigate compositional heterogeneity between taxa, but this comes at the cost of information loss, and the reliability of such methods is so far unproven. Here we demonstrate that site-exclusion methods produce erratic phylogenetic estimates of mitochondrial origin. Thus, previous phylogenetic hypotheses on the origin of mitochondria based on pretreated datasets should be re-evaluated. We applied alternative strategies to reduce phylogenetic noise by systematic taxon sampling while keeping site substitution information intact. Cross-validation based on a series of trees placed mitochondria robustly within Alphaproteobacteria, sharing an ancient common ancestor with Rickettsiales and currently unclassified marine lineages.

The origin of mitochondria is one of the defining events in the history of life. Gene-network analyses[1–4] and marker gene-based phylogenomic inference have generally reached a consensus that mitochondria have an alphaproteobacterial common ancestor[5], yet the specific relationship of mitochondria to alphaproteobacterial taxa remains an important evolutionary issue. Phylogenetic placement of mitochondria within the tree of Alphaproteobacteria has been hampered by variation in mitochondrial DNA nucleotide composition and substitution rates as well as strong phylogenetic artefacts associating mitochondria with some fast-evolving alphaproteobacterial lineages such as Rickettsiales and Pelagibacterales, resulting in erroneous branching patterns (Supplementary Note 1). To minimize the possible influence of long-branch attraction coupled with convergent compositional signals, various strategies have been applied, such as the use of nucleus-encoded mitochondrial genes[4,6,7], site or gene exclusion[8–10], protein recoding[10] and the use of heterogeneity-tolerant models[6,11] (Supplementary Note 2). These attempts have not converged but have instead generated contradictory hypotheses including (1) mitochondria root in or are sisters of Rickettsiales[7,12], which are all obligate endosymbionts (but see ref. [13]); (2) mitochondria are sisters of free-living Alphaproteobacteria such as *Rhodospirillum rubrum*[9], Rhizobiales and Rhodobacterales[4]; (3) mitochondria are neighbours to a group of uncultured marine bacteria[14] and (4) mitochondria are most closely related to the most abundant marine surface Alphaproteobacteria, SAR11 (referred to as Pelagibacterales in this study)[15,16]. The first hypothesis has been reported most frequently so far, while the last has been identified by several independent groups as being the result of compositional convergence artefacts[8,11,14]. Recently, Martijn et al.[17] revisited the topic, reporting that when compositional heterogeneity of the protein sequence alignments was reduced by excluding sites from the amino acid alignment, the entire alphaproteobacterial class formed a sister group to mitochondria[17]. Their conclusion is at odds with the long-standing phylogenetic consensus that mitochondria originated within Alphaproteobacteria[18]. While excluding compositionally heterogenous sites might reduce phylogenetic noise and mitigate systematic errors, it will also necessarily lead to loss of phylogenetic information (Supplementary Note 3). At what point does the exclusion of sites exclude signals of the true evolutionary connection between mitochondria and Alphaproteobacteria? This question was not adequately addressed in their publication, and a similar concern has been voiced by Gawryluk[19]. To explore this important evolutionary issue, we systematically examine the impact of site-exclusion methods on the phylogenetic affiliations of mitochondria to Alphaproteobacteria. The results uncover hitherto unrecognized pitfalls of site-exclusion approaches. Subsequent taxon sampling and verification approaches robustly place mitochondria within Alphaproteobacteria.

Results

We used several approaches to systematically investigate the effects of data exclusion on determining phylogenetic relationships between mitochondria, Alphaproteobacteria and outgroups. First, different site-exclusion methods were investigated by cross-validation to see if alternative trends in tree topological change were observed. Specifically, five metrics with different

[1]Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Department of Ocean Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China. [2]Academy for Advanced Interdisciplinary Studies, Southern University of Science and Technology (SUSTech), Shenzhen, China. [3]Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, China. [4]Department of Bioinformatics, Putuo People's Hospital, Tongji University, Shanghai, China. [5]Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy. [6]Institute of Molecular Evolution, Heinrich Heine University, Düsseldorf, Germany. [7]These authors contributed equally: Lu Fan, Dingfeng Wu, Vadim Goremykin.
✉e-mail: fanl@sustech.edu.cn; bill@hhu.de; rxzhu@tongji.edu.cn

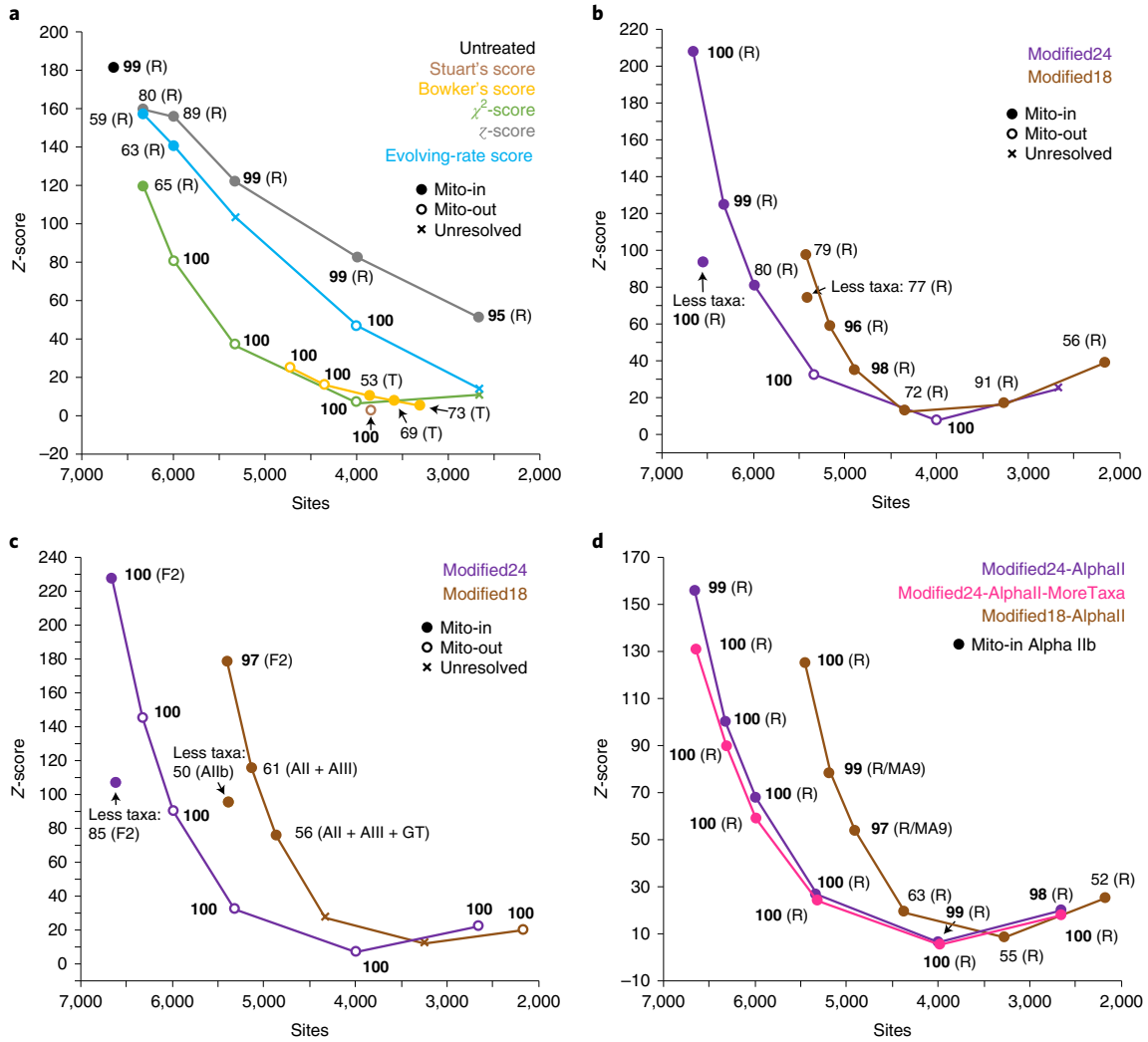

Fig. 1 | Relationships between alignment sites, the phylogenetic position of mitochondria and model fit (mean square heterogeneity across taxa test) based on different datasets, site-exclusion approaches and taxon-selection approaches. a–d, Bayesian inference with model CAT+GTR was conducted. The *x* axis shows the number of sites for phylogenetic inference in each dataset. The *y* axis shows the *Z*-scores of the 'mean square heterogeneity across taxa' posterior predictive test. Numbers beside markers show node support values (posterior probability support values) of the consensus trees. Values ≥95 are in bold. Labels in parentheses show the closest relatives of mitochondria in the tree (R, Rickettsiales; T, *T. mobilis*; F2, FEMAG II; AII, Alpha II; AIII, Alpha III; GT, *Geminicoccus roseus* and *T. mobilis*; MA9, MarineAlpha9). Mito-in (filled circles) means mitochondria branch within Alphaproteobacteria; mito-out (open circles) means mitochondria branch outside Alphaproteobacteria and mito-in Alpha IIb means mitochondria branch within the Alpha IIb clade of Alphaproteobacteria. **a**, Site-exclusion methods applied to the '24-alphamitoCOGs' dataset[17]. Trees are shown in Supplementary Figs. 1–22. **b**, χ^2-score-based site-exclusion and taxon-reduction methods applied to subsets of the 'Modified24' and the 'Modified18' datasets containing only the backbone, Rickettsiales and mitochondrial sequences. Trees are shown in Supplementary Figs. 35 and 37–42. **c**, χ^2-score-based site-exclusion and taxon-reduction methods applied to subsets of the 'Modified24' and the 'Modified18' datasets containing only the backbone, FEMAGs and mitochondrial sequences. Trees are shown in Supplementary Figs. 36 and 42–48. **d**, χ^2-score-based site-exclusion method applied to the subsets of the 'Modified24-AlphaII', 'Modified24-AlphaII-MoreTaxa' and 'Modified18-AlphaII' datasets. Trees are shown in Fig. 4 and Supplementary Figs. 50–58.

principles—Stuart's score, Bowker's score, χ^2-score, ζ-score and evolving-rate score—were implemented (Supplementary Table 1). Site-excluded subsets of the '24-alphamitoCOGs' dataset[17] were generated using the five methods with a series of cut-off values (Supplementary Table 2). Bayesian trees were reconstructed, and posterior predictive tests of model fit were conducted following the same procedures reported[17] (Supplementary Figs. 1–22). In general, all site-exclusion methods led to a decrease in amino acid compositional heterogeneity between taxa and improvement in fit between the data and the model, as demonstrated by reduced *Z*-scores of the maximum/mean square compositional heterogeneity tests[20] (Fig. 1a, Supplementary Fig. 23 and Supplementary Table 2). The more sites excluded, the greater the improvement. When the same number of sites were excluded, Stuart's score, Bowker's score and χ^2-score methods were more efficient in improving model fit than were ζ-score and evolving-rate-score methods. We found two notable cases of data irreproducibility in comparison to the original publication[12] (Supplementary Note 4) and observed a consistent trend of tree topology shift for the Stuart's score and χ^2-score methods (Fig. 1). Specifically, when 10% or more of the sites were excluded, the connection between mitochondria and Rickettsiales was broken and an Alphaproteobacteria-sister topology (referred to here as 'mito-out', in contrast with 'mito-in', where mitochondria are within Alphaproteobacteria) with strong node support was

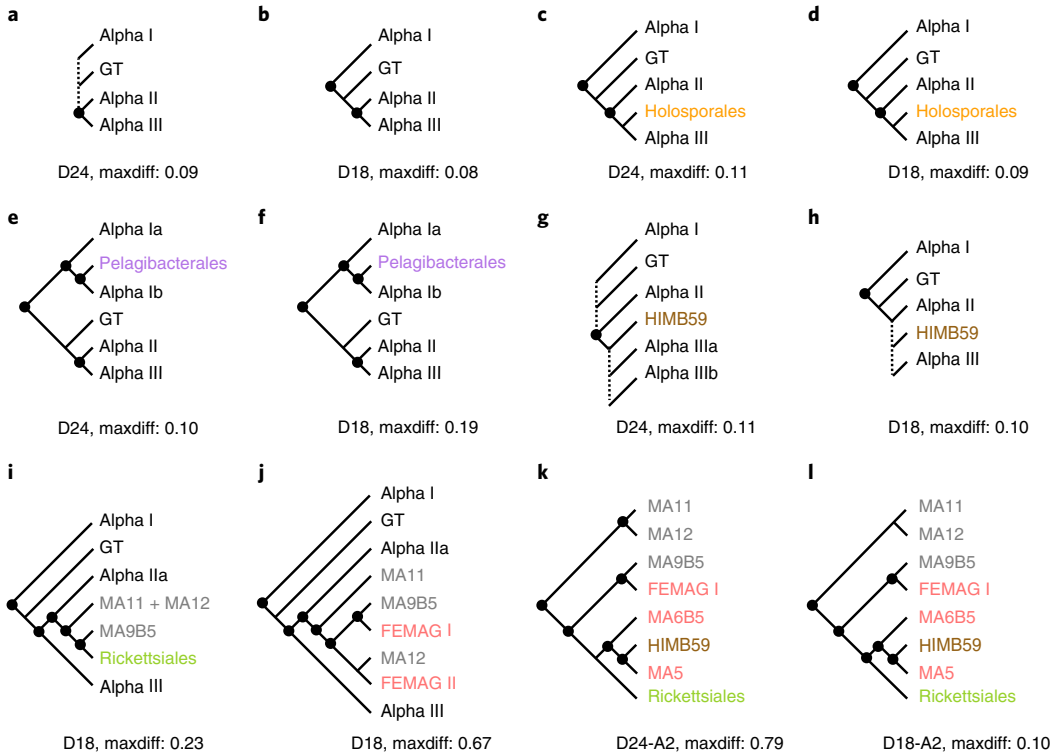

Fig. 2 | Schematic phylogenetic relationships of alphaproteobacterial subgroups. Different alphaproteobacterial subgroups are shown in different colours and named according to Supplementary Table 3. Taxa and taxonomic subgroups shown in black are the backbone taxa. Taxa shown in grey are backbone taxa belonging to the Alpha IIb subclade. Dots demonstrate node posterior probability support values ≥95%. **a**–**j**, Schematic drawings of Bayesian trees in Supplementary Figs. 25a (**a**), 25b (**b**), 26a (**c**), 26b (**d**), 27a (**e**), 27b (**f**), 28a (**g**), 28b (**h**), 29b (**i**) and 30b (**j**). Trees are rooted. Outgroup taxa and *Magnetococcus marinus* MC-1 are not shown. No schematic trees were drawn for Supplementary Figs. 29a and 30a as their maxdiff values are equal to one. Labels below each panel indicate the dataset and the maxdiff value for the tree. D24, the 'Modified24' dataset; D18, the 'Modified18' dataset. **k**,**l**, Schematic drawings of Bayesian trees in Supplementary Fig. 49. Trees are rooted and outgroup taxa are not shown. D24-A2, the 'Modified24-AlphaII' dataset; D18-A2, the 'Modified18-AlphaII' dataset.

established (Fig. 1a, Supplementary Figs. 2–7 and 23). A similar tendency to lose the mitochondria–Rickettsiales connection was also observed using the evolving-rate score method when 20% or more of the sites were excluded (Fig. 1a, Supplementary Figs. 8–12 and 23). However, there were three cases where the position of mitochondria was unresolved in the χ^2-score and evolving-rate score methods, suggesting that the mito-out topology is likely unstable and vulnerable to site-exclusion effects. Notably, along with the improved model fit, all five trees based on the ζ-score method supported the Rickettsiales-sister topology, including several cases of node support values >95% (Fig. 1a, Supplementary Figs. 13–17 and 23). The Bowker's score method generated trees supporting either the mito-out topology or with mitochondria branching in proximity to *Tistrella mobilis* (Fig. 1a, Supplementary Figs. 18–22 and 23).

The main finding of our extensive validation of site-exclusion methods is that the shift from Rickettsiales-sister to the mito-out topology produced by reducing the number of sites is purely method dependent. Excluding different sites by using different methods produces a consistent tendency to improve fit between model and data but inconsistent phylogenetic positions of mitochondria. Thus, tree topology and the compositional heterogeneity of the '24-alphamitoCOGs' dataset are not causally correlated. Moreover, the possibility that historical signals are lost during site exclusion and that, consequently, the reported mito-out topology[17] is a long-branch attraction effect introduced by the distant outgroup cannot be ruled out. Indeed, Martijn et al.[17] failed to exclude long-branch attraction effects between mitochondria and the distant outgroup, as detailed in Supplementary Note 3.

Large datasets including many fast-evolving lineages, such those reported by Martijn et al.[17], may exacerbate systematic errors in phylogeny. Therefore, we applied a strategy exploring the phylogenetic relationships of mitochondria with individual alphaproteobacterial lineages. First, the '24-alphamitoCOGs' dataset[17] was modified, and two master datasets, 'Modified24' and 'Modified18', were generated. In the latter dataset, mitochondrial and Rickettsiales sequences were replaced by alternatives with high genomic GC contents (Supplementary Fig. 24, Supplementary Tables 3 and 4, and Supplementary Note 5). Taxa in each dataset were classified into 'slow-evolving' and 'fast-evolving' clades, generally based on their genomic GC contents (Supplementary Table 3), and Bayesian trees were reconstructed using combinations of these clades (Supplementary Table 5 and Supplementary Figs. 25–36).

The tree topology of Alphaproteobacteria themselves is an issue in its own right[21]. When fast-evolving taxa were excluded, Alphaproteobacteria could be classified into four major clades: Alpha I, Alpha II, Alpha III and GT (Fig. 2a,b, Supplementary Table 3 and Supplementary Note 6). We assigned these slow-evolving Alphaproteobacteria as the 'backbone' taxa to which each of the five fast-evolving subgroups was added. Notably, taxon addition preserved a topology in which all four backbone clades maintain their monophyly. The 'Modified24' and the 'Modified18' datasets produced robust and consistent phylogenetic positions for Holosporales and Pelagibacterales. Specifically, Holosporales were placed adjacent to the entire clade of Alpha III (Fig. 2c,d), and Pelagibacterales branched as the sister of Alpha Ib (Fig. 2e,f). The exact phylogenetic position of alpha proteobacterium HIMB59 could not be resolved.

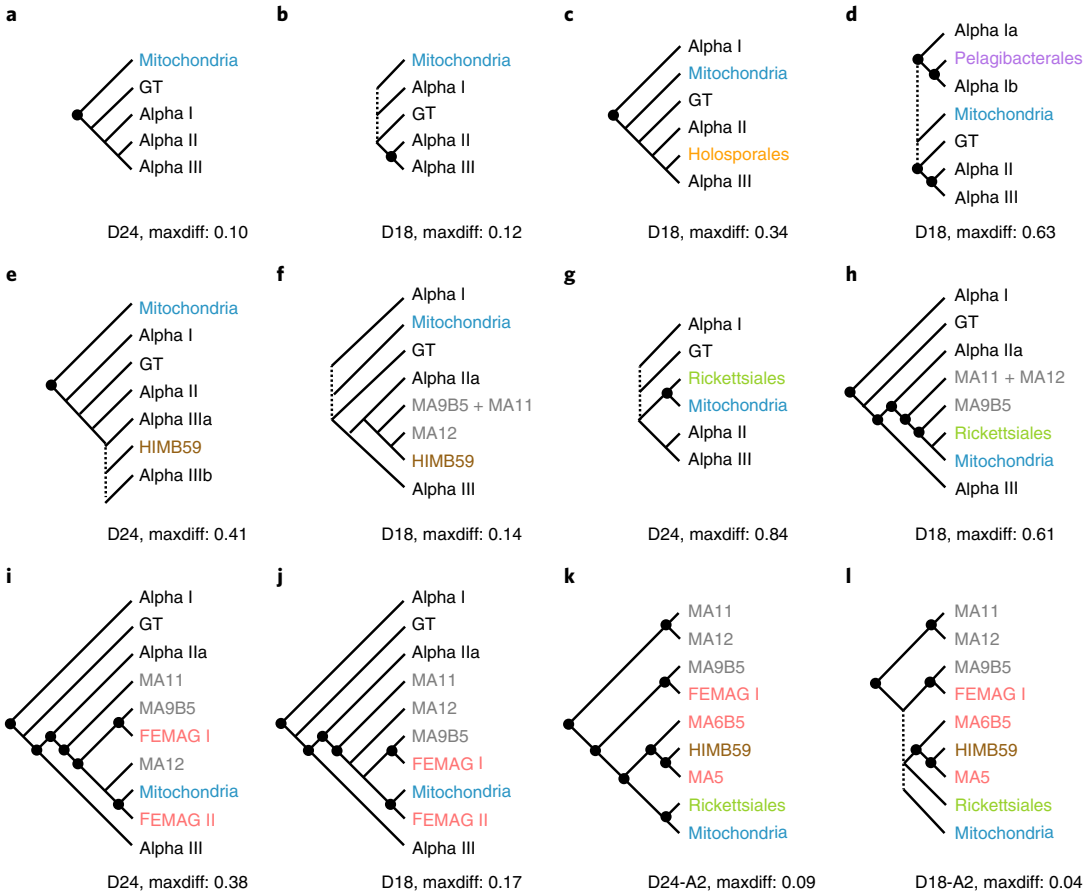

Fig. 3 | Schematic phylogenetic relationships of mitochondria and alphaproteobacterial subgroups. Different alphaproteobacterial subgroups are shown in different colours and named according to Supplementary Table 3. Taxa and taxonomic subgroups shown in black represent the backbone taxa. Taxa shown in grey are backbone taxa belonging to the Alpha IIb subclade. Dots demonstrate node posterior probability support values ≥95%. **a**–**j**, Schematic drawings of Bayesian trees in Supplementary Figs. 31a (**a**), 31b (**b**), 32b (**c**), 33b (**d**), 34a (**e**), 34b (**f**), 35a (**g**), 35b (**h**), 36a (**i**) and 36b (**j**). Trees are rooted. Outgroup taxa and *M. marinus* MC-1 are not shown. No schematic trees were drawn for Supplementary Figs. 32a and 33a as their maxdiff values are equal to one. Labels below each panel indicate the dataset and the maxdiff value for the tree. D24, the 'Modified24' dataset; D18, the 'Modified18' dataset. **k**,**l**, Schematic drawings of the Bayesian trees in Supplementary Fig. 50. Trees are rooted and outgroup taxa are not shown. D24-A2, the 'Modified24-AlphaII' dataset; D18-A2, the 'Modified18-AlphaII' dataset.

It was placed adjacent to Alpha II and Alpha III, suggesting a weak phylogenetic connection to these two clades (Fig. 2g,h). Trees of the subsets of the 'Modified24' dataset containing either Rickettsiales or fast-evolving metagenome-assembled genomes (FEMAGs) failed to converge (maxdiff value of 1.0, sample size 30,000) (Supplementary Figs. 29–30a and Supplementary Table 5). Based on the 'Modified18' dataset, Rickettsiales and FEMAG I showed a strong connection to the backbone marine alphaproteobacterium MarineAlpha9 Bin5 (Fig. 2i,j), while FEMAG II was linked to MarineAlpha12 Bin1. We compare our results to previous studies in Supplementary Note 7.

We then added mitochondria to the trees of the backbone taxa, solely or in combination with fast-evolving clades of Alphaproteobacteria. Mitochondria by themselves were placed outside of backbone taxa; otherwise, their phylogenetic relationship to the backbone clades was not resolved (Fig. 3a,b). Similar results were observed in trees that included mitochondria in combination with Holosporales, Pelagibacterales or alpha proteobacterium HIMB59, suggesting the backbone taxa and these fast-evolving lineages have little evolutionary affinity to mitochondria (Fig. 3c–f). Notably, when mitochondria were present, alpha proteobacterium HIMB59 was placed in the Alpha IIb clade based on the 'Modified18' dataset (Fig. 3f). By contrast, an apparent phylogenetic connection of mitochondria to Rickettsiales and FEMAG II was observed in both datasets (Fig. 3g–j). Specifically, mitochondria and Rickettsiales either grouped together inside the Alpha IIb clade or remained separate from the four backbone clades, while mitochondria and FEMAG II formed sisters within Alpha IIb. To test the potential impact of model violation on these results, χ^2-score-based site-exclusion and taxon-reduction approaches were conducted on the datasets that contained mitochondria along with either Rickettsiales or FEMAGs (Fig. 1b,c, Supplementary Table 5 and Supplementary Figs. 23, 37–48). Model violation was successfully alleviated. The phylogenetic connection between mitochondria and Rickettsiales was preserved when using the 'Modified18' dataset but was lost when 20% or more sites were excluded when using the 'Modified24' dataset. In comparison, the phylogenetic connection between mitochondria and FEMAG II was lost when 5% or more sites were excluded when using either dataset. Importantly, the mito-in topology was preserved in all the reduced-taxon approaches. These results suggest that the strong connection between mitochondria and Rickettsiales is likely one of the true historical signals that persist when model violation is minimized.

As Rickettsiales, alpha proteobacterium HIMB59, FEMAG I and FEMAG II all individually showed phylogenetic connections to backbone taxa of Alpha IIb, evolutionary relationships between these lineages were then specifically investigated by setting Alpha

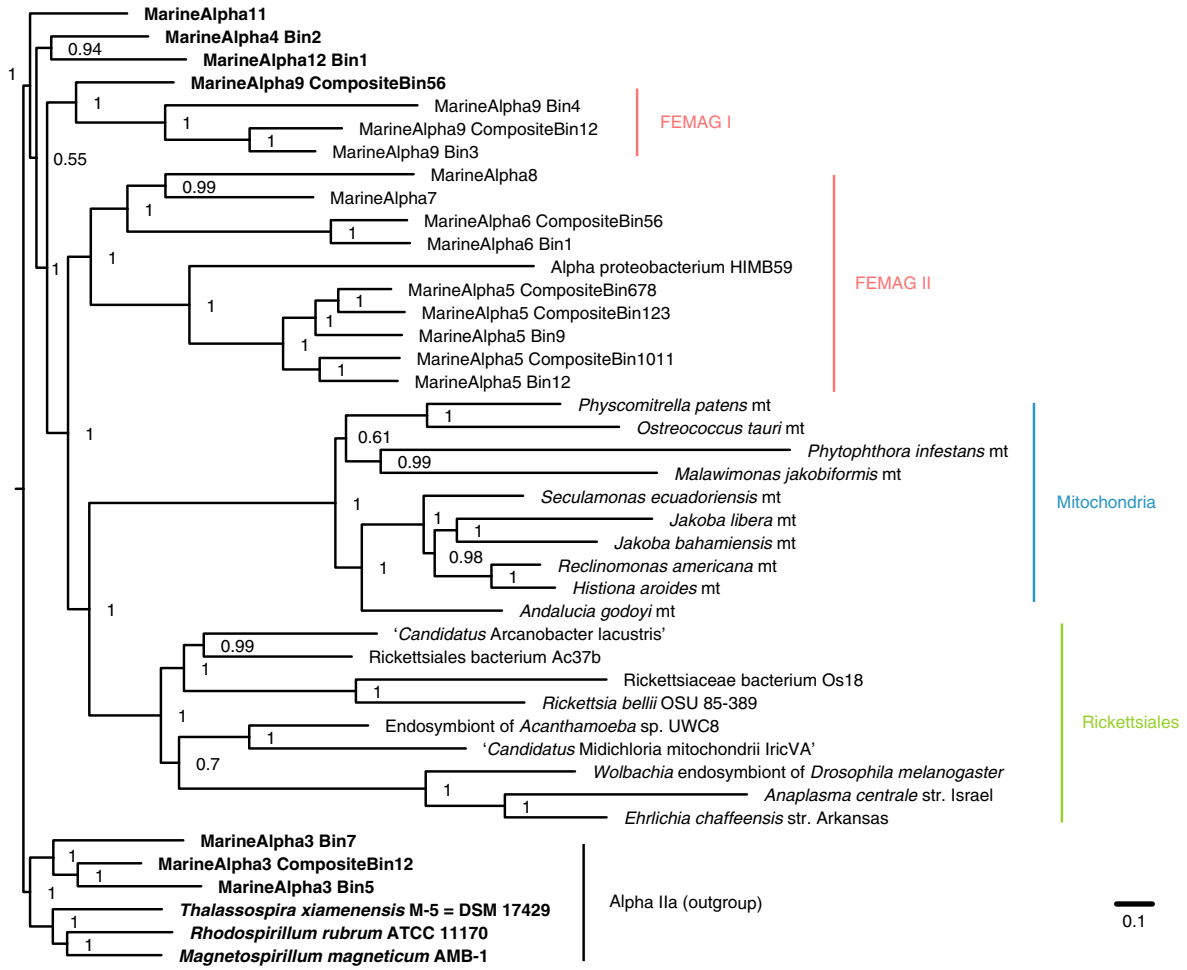

Fig. 4 | Phylogenetic relationships of mitochondria and Alpha IIb Alphaproteobacteria. Bayesian inference was conducted with model CAT+GTR. Numbers beside nodes show posterior probability support values. The tree is rooted to the outgroup Alpha IIa. Alphaproteobacterial subgroup names are coloured according to Fig. 3. Mitochondrial genomes are indicated by 'mt'. Taxa with names in bold are backbone Alpha II taxa (Supplementary Table 3).

IIa as the outgroup (Supplementary Figs. 49 and 50). Because Alpha IIa taxa are closely related to Alpha IIb taxa, we were able to obtain trees that were little impacted by outgroup attraction. This generated two datasets, 'Modified24-AlphaII' and 'Modified18-AlphaII', that produced consistent results. MarineAlpha11 and MarineAlpha12 formed a monophyletic clade (Fig. 2k,l) and MarineAlpha9 Bin5/CompositeBin56 grouped with FEMAG I. Alpha proteobacterium HIMB59 robustly branched within FEMAG II as the sister of MarineAlpha5 bins. Rickettsiales was in a sister relationship with FEMAG II. When mitochondria were present, the topology of all Alphaproteobacteria was preserved when using either dataset (Fig. 3k,l). Mitochondria were placed as the sister of Rickettsiales with strong node support when using the 'Modified24-AlphaII' dataset. For the 'Modified18-AlphaII' dataset, mitochondria were adjacent to two clades, with unresolved relationships to those clades. The first clade consists of FEMAG II, alpha proteobacterium HIMB59 and Rickettsiales, while the other consists of FEMAG I and MarineAlpha9 Bin5. Despite the relationships being unresolved, the placement of mitochondria within Alpha IIb was robust. Our results suggest that Rickettsiales may be the closest relative of mitochondria and that Rickettsiales and mitochondria share a common ancestor with certain extant marine planktonic Alphaproteobacteria.

To verify the potential impact of model violation on this topology, we first generated an additional dataset based on 'Modified24-AlphaII' (Fig. 4 and Supplementary Table 3) with more taxa and then conducted a χ^2-score-based site-exclusion treatment on the three above-mentioned Alpha II taxa datasets (Fig. 1d and Supplementary Figs. 51–58). All treatments produced trees that placed mitochondria robustly within the Alpha IIb clade. Specifically, the Rickettsiales-sister topology is preserved except in two cases where the relationship between mitochondria, Rickettsiales and MarineAlpha9 bins was unresolved. Therefore, the placement of mitochondria with Rickettsiales and within Alpha IIb is not an error of model violation. Instead, it has components of true historical signals as detected by the present data.

Discussion

We have demonstrated that the Alphaproteobacteria-sister, or mito-out, topology of mitochondria reported by Martijn et al.[17] is the result of the loss of critical evolutionary signals via arbitrary site-exclusion procedures. By using additional taxa, multiple datasets, alternative site-exclusion methods, taxon sampling strategies and model violation justification, we successfully verified the evolutionary connection between mitochondria and alphaproteobacterial lineages including Rickettsiales and several marine planktons derived from metagenomes. This tree topology is robust to various parameters and is unlikely to be a result of phylogenetic artefacts, as indicated by several lines of evidence (Supplementary Note 8).

Our result is in agreement with numerous previous studies that suggested a phylogenetic connection between mitochondria and Rickettsiales[5]. Future work (including both metagenomic and cultivation efforts) on the subclade II of Alphaproteobacteria, especially Alphaproteobacteria from marine sediments or other microoxic environments, should provide further insights into the closest extant relatives of mitochondria and possible hints to the metabolic nature of the common ancestor of mitochondria (Supplementary Note 9).

Methods

Implementation of site-exclusion metrics. To obtain the '24-alphamitoCOGs' dataset[17], the file 'alphaproteobacteria_mitochondria_untreated.aln' was downloaded from https://doi.org/10.5061/dryad.068d0d0. As the names of some MarineAlpha bins in this file are not consistent with the phylogenetic trees in the original paper, we obtained the name mapping file from J. Martijn on 4 July 2018. In this dataset, χ^2-score-based site exclusion was achieved by applying an equation introduced previously[8]. ζ-score is a metric specifically designed to cope with strong GC-content-related amino acid compositional heterogeneity in datasets of alphaproteobacterial phylogeny[21]. ζ-scores of sites were calculated according to a previously reported method[21]. A method implemented in IQ-TREE for fast-evolving site selection was also included for comparison, since long-branch attraction caused by fast-evolving species in Alphaproteobacteria and mitochondria is a potential issue[22]. The evolving-rate score method was based on conditional mean site rates estimated under the LG+C60+F+R6 model set in IQ-TREE (v1.5.5) using the '-wsr' flag[22]. Sites were sorted based on these three metrics, and the top 5%, 10%, 20%, 40% and 60% were excluded from downstream phylogenetic analyses (Supplementary Table 2). Stuart's test and Bowker's test are two typical evaluation metrics of symmetry violation[23]. The site-excluded dataset ('alphaproteobacteria_mitochondria_stationarytrimmed.aln') based on Stuart's score was downloaded from https://doi.org/10.5061/dryad.068d0d0, and that based on Bowker's score is described in detail in Supplementary Methods.

Phylogenetic inference and model violation tests. Site-excluded protein sequence alignments were treated using trimAl '-gappyout' (v1.4)[24] before phylogenetic inference. Bayesian trees were produced using PhyloBayes MPI (v1.8)[25]. The CAT+GTR models were used. As non-parametric models were used, a priori was not specified. Four chains were run for each consensus tree, and for each chain over 15,000 cycles (5,000 burn-in) were conducted, until a maxdiff value lower than 0.3 was reached. Otherwise, non-converged chains were continually run to over 20,000 cycles (Supplementary Tables 2 and 5). Posterior predictive tests were conducted using PhyloBayes MPI with the 'readpb_mpi -x 5000 50 -allppred' command.

Genome and marker protein selection of the 'Modified18' dataset. The 'Modified18' dataset is modified from the '24-alphamitoCOGs' dataset[17]. Specifically, since composite bins contain sequences from multiple naturally existing genomes, we extracted the bin with the highest marker protein coverage from each composite bin (Supplementary Table 3) to minimize possible assembly-induced artefacts. Five less-AT-rich mitochondria (GC content 45.1–52.2%) and five less-AT-rich Rickettsiales (GC content 38.2–49.8%) were selected to replace the mitochondrial and Rickettsiales groups in the original dataset (Supplementary Table 3). The GC-poor versus GC-rich amino acid (FYMINK/GARP) marker protein ratios of the reselected mitochondria and Rickettsiales ranged from 0.955 to 1.329 and from 1.013 to 2.330, respectively (Supplementary Fig. 24). All relevant genomes were downloaded from the RefSeq database of NCBI on 21 July 2018.

For quality control of the 24 marker proteins of the original dataset, sequences of these proteins were downloaded from the MitoCOGs[26] database and then aligned using MAFFT-L-INS-I (v7.055b)[27]. Alignment of each protein was trimmed using trimAl '-gappyout' (v1.4)[24]. Protein-specific *e*-values were determined with distributions of positive and negative sequences. For each protein, sequences classified as proteins in the MitoCOGs database were chosen as positive controls and sequences classified as other proteins were chosen as negative controls. The *e*-value distribution of the positive and negative sequences was calculated using Hmmer (v3.2.1)[28]. The protein-specific *e*-values of the positive sequences were the minimum of the 95% quantile *e*-values of those sequences, while the *e*-values of the negative sequences were the minimum of those negative sequences. We searched these 24 proteins individually in the genomes using Hmmer based on protein-specific *e*-values of the HMM models. The obtained proteins were processed for ML tree reconstruction using IQ-TREE under the LG+C60+F model. Copies identified as paralogs, possible contaminants or events of lateral gene transfer in each gene tree were removed. '*Candidatus* Paracaedibacter symbiosus' was excluded as multiple contaminant proteins were detected in its genome, and we think its genome likely suffers from assembly contamination. MitoCOG0003 and MitoCOG0133 were excluded as they were detected in few genomes. MitoCOG00052, MitoCOG00060, MitoCOG00066 and MitoCOG00071 were excluded as they were absent in reselected mitochondrial genomes. Consequently, 18 marker proteins were selected (Supplementary Table 4). Except for outgroup species including Betaproteobacteria, Gammaproteobacteria and Magnetococcales, genomes containing 16 or more of the 18 marker proteins were kept. Furthermore, we removed some redundant MarineAlpha bins of the original dataset based on pairwise similarity of marker proteins using BLASTP (v2.6.0+, identity ≥0.99 and coverage ≥0.95). As a result, 61 genomes were kept for downstream analysis. Before phylogenetic inference, selected proteins were aligned using MAFFT-L-INS-i. Low-quality columns were removed by BMGE '-m BLOSUM30' (v1.12)[29] and the multiple sequence alignments after quality control were concatenated.

Taxa selection of the other datasets. The 'Modified24' dataset is modified based on the '24-alphamitoCOGs' dataset. Taxa corresponding to those in the 'Modified18' dataset were included, except Rickettsiales and mitochondria. Instead, the Rickettsiales and mitochondrial sequences in the '24-alphamitoCOGs' dataset were used. Taxonomic subsets of the 'Modified24' and the 'Modified18' datasets, shown in Figs. 2 and 3, were generated according to the taxonomic groups shown in Supplementary Table 3. The subsets of 'Modified24' and 'Modified18' with Rickettsiales or FEMAGs but fewer taxa shown in Fig. 1b,c and Supplementary Figs. 42 and 48 were generated by removing taxa with high *Z*-scores of heterogeneity, as shown in Supplementary Table 6. The datasets 'Modified24-AlphaII', 'Modified24-AlphaII-MoreTaxa', 'Modified18-AlphaII', shown in Fig. 1d and Supplementary Table 3, were generated by selecting taxa belonging to or closely related to the backbone Alpha IIb lineages shown in Figs. 2 and 3. χ^2-score-based site-exclusion approaches for the subsets of 'Modified24' and 'Modified18' with Rickettsiales or FEMAGs, as shown in Fig. 1b,c, and for 'Modified24-AlphaII', 'Modified24-AlphaII-MoreTaxa' and 'Modified18-AlphaII', as shown in Fig. 1d, were conducted by applying an equation introduced previously[8].

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

Data availability

The alignments and tree files generated in this study have been deposited in figshare (https://doi.org/10.6084/m9.figshare.12347216) (ref. [30]).

Code availability

The script of the Bowker's test score-based site-exclusion method is available as Supplementary Software.

Received: 7 September 2019; Accepted: 29 May 2020;
Published online: 13 July 2020

References

1. Ku, C. et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).
2. 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).
3. Abhishek, A., Bavishi, A., Bavishi, A. & Choudhary, M. Bacterial genome chimaerism and the origin of mitochondria. *Can. J. Microbiol.* **57**, 49–61 (2011).
4. Atteia, A. et al. A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. *Mol. Biol. Evol.* **26**, 1533–1548 (2009).
5. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
6. Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol. Biol. Evol.* **29**, 1277–1289 (2012).
7. Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Sci. Rep.* **5**, 7949 (2015).
8. Viklund, J., Ettema, T. J. & Andersson, S. G. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
9. Esser, C. et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643–1660 (2004).
10. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.* **23**, 74–85 (2006).
11. 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).

12. Viale, A. M. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic organelles. *FEBS Lett.* **341**, 146–151 (1994).
13. Castelli, M. et al. *Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of Rickettsiales. *ISME J.* **13**, 2280–2294 (2019).
14. Brindefalk, B., Ettema, T. J., Viklund, J., Thollesson, M. & Andersson, S. G. A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade. *PLoS ONE* **6**, e24457 (2011).
15. Thrash, J. C. et al. Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.* **1**, 13 (2011).
16. 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).
17. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
18. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283**, 1476–1481 (1999).
19. Gawryluk, R. M. R. Evolutionary biology: A new home for the powerhouse? *Curr. Biol.* **28**, R798–R800 (2018).
20. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).
21. 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).
22. 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).
23. Jermiin, L. S., Jayaswal, V., Ababneh, F. M. & Robinson, J. Identifying optimal models of evolution. *Methods Mol. Biol.* **1525**, 379–420 (2017).
24. 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).
25. 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).
26. 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).
27. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
28. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
29. 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).
30. Fan, L. et al. Mitochondria and Alphaproteobacteria phylogenetic study alignments and tree files. *figshare* https://doi.org/10.6084/m9.figshare.12347216 (2020).

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (91851210, 91951120, 41530105 and 81774152), the European Research Council (ERC 666053), the Shenzhen Key Laboratory of Marine Archaea Geo-Omics, Southern University of Science and Technology (ZDSYS201802081843490), the Shenzhen Science and Technology Innovation Commission (JCYJ20180305123458107), the Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (K19313901) and the VW foundation (93 046). Computation in this study was supported by the Centre for Computational Science and Engineering at the Southern University of Science and Technology.

Author contributions

L.F., W.F.M. and R.Z. conceived this study. L.F., D.W., V.G., J.X., Y.X. and S.G. were involved in data analysis. L.F., V.G., C.Z., W.F.M. and R.Z. interpreted the results and drafted the manuscript. All authors participated in the critical revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-1239-x.

Correspondence and requests for materials should be addressed to L.F., W.F.M. or R.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Lu Fan, William F. Martin, Ruixin Zhu

Last updated by author(s): May 21, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see 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	No software was used.
Data analysis	IQ-TREE (v1.5.5 & v1.6.12); trimAl (v1.4); PhyloBayes MPI (v1.8); MAFFT-L-INS-I (v7.055b); Hmmer (v3.2.1); BLASTP (v2.6.0+); BMGE (v1.12); CheckM (v1.0.11); KofamKOALA (v1.2.0); MUSCLE (v3.8); the script as Supplementary Software.

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 Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The alignments and tree files generated in this study has been deposited in figshare (DOI: 10.6084/m9.figshare.12347216) (ref. 30).

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

Ecological, evolutionary & environmental sciences study design

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

Study description	This is a study of phylogeny. Consensus phylogenetic trees were compared one to other manually. Statistics appled to node bootstrapping and model violation tests based on multiple trees by following standard program setup (1000 iterations for ML trees and numbers specified in Supplementary Tables for Bayesian trees).
Research sample	N.A.
Sampling strategy	N.A.
Data collection	Protein sequences and genome sequences were all downloaded from open resources. The '24-alphamitoCOGs' dataset in Martijn et al. (2018), file 'alphaproteobacteria_mitochondria_untreated.aln' was downloaded from https://datadryad.org//resource/doi:10.5061/dryad.068d0d0. The site-excluded dataset ('alphaproteobacteria_mitochondria_stationarytrimmed.aln') based on Stuart's test score was downloaded from https://datadryad.org//resource/doi:10.5061/dryad.068d0d0. All relevant genomes were downloaded from the RefSeq database of NCBI on 21 July 2018.
Timing and spatial scale	N.A.
Data exclusions	No data were excluded.
Reproducibility	N.A.
Randomization	N.A.
Blinding	N.A.

Did the study involve field work? ☐ Yes ☒ No

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