

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

Sergio Munoz-Gomez (✉ smunozgo@asu.edu)

Arizona State University

Edward Susko

Dalhousie University

Kelsey Williamson

Dalhousie University

Laura Eme

Université Paris-Saclay

Claudio Slamovits

Dalhousie University

David Moreira

Université Paris-Saclay

Purificacion Lopez-Garcia

Université Paris-Saclay

Andrew Roger

Dalhousie University <https://orcid.org/0000-0003-1370-9820>

Article

Keywords: Alphaproteobacteria, mitochondria, evolutionary genetics

Posted Date: May 28th, 2021

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

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Ecology & Evolution on January 13th, 2022. See the published version at <https://doi.org/10.1038/s41559-021-01638-2>.

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

3 **Authors:** Sergio A. Muñoz-Gómez^{1,4*}, Edward Susko², Kelsey Williamson¹, Laura Eme³, Claudio H.
4 Slamovits¹, David Moreira³, Purificación López-García³, and Andrew J. Roger^{1*}

5 **Affiliations:**

6 ¹Centre for Comparative Genomics and Evolutionary Bioinformatics, Department of Biochemistry and
7 Molecular Biology, Dalhousie University, Halifax, Canada.

8 ²Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada.

9 ³Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France.

10 ⁴Current affiliation: Ecologie Systématique Evolution, Université Paris-Saclay, AgroParisTech, Orsay,
11 France.

12 *Correspondence to: sergio.munoz@universite-paris-saclay.fr and andrew.roger@dal.ca

13 **Abstract**

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

27 **Introduction**

28 Mitochondria stem from an ancient endosymbiosis that occurred during the origin of eukaryotic cells¹. As
29 a result, all extant eukaryotes have mitochondria or evolved from mitochondrion-bearing ancestors¹⁻³.
30 Some hypotheses have it that mitochondria provided excess energy required for the origin of eukaryotic
31 complexity⁴, whereas others suggest that mitochondrial symbiosis brought efficient aerobic respiration
32 into a more complex proto-eukaryote⁵. The nucleocytoplasm of eukaryotes is now known to be most
33 closely related to Asgard archaea⁶⁻⁸. Mitochondria, on the other hand, have been known for decades to
34 be phylogenetically associated with the *Alphaproteobacteria*^{9,10,11}. However, the precise relationship
35 between mitochondria and the *Alphaproteobacteria*, or any of its sub-groups, has been elusive and
36 remains a matter of intense debate (e.g., see ^{11,12}). Settling this debate will provide insights into the nature
37 of the mitochondrial ancestor and the ecological setting of its endosymbiosis with the host cell¹.

38 Mitochondria have been placed in various regions of the tree of the *Alphaproteobacteria*. Most early
39 studies suggested that mitochondria were most closely related to the *Rickettsiales*¹³⁻²⁰ (*Rickettsiales*-
40 sister hypothesis), a group classically known for comprising intracellular parasites. This led many to
41 believe that mitochondria evolved from parasitic alphaproteobacteria^{18,21}. However, relationships between
42 mitochondria and the *Pelagibacterales*^{22,23}, *Rhizobiales*²⁴, or *Rhodospirillales*²⁵ have also been proposed.
43 These alternative proposals suggested that mitochondria may have evolved from either streamlined or
44 metabolically versatile free-living alphaproteobacteria²²⁻²⁵. Most recently, the phylogenetic placement of
45 mitochondria has been vividly debated^{11,12}. One study found mitochondria as a sister group to the entire
46 *Alphaproteobacteria* (i.e., the *Alphaproteobacteria*-sister hypothesis)¹¹. This conclusion was supported by
47 the inclusion of novel alphaproteobacterial metagenome-assembled genomes (MAGs) from worldwide

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

53 There are several reasons why it is difficult to confidently place mitochondria among their
54 alphaproteobacterial relatives. First, the evolutionary divergence between mitochondria and their closest
55 bacterial relatives is estimated to have occurred >1.5 billion years ago^{27,28}. This has erased the historical
56 signal (e.g., through multiple amino acid replacements) that was originally present in the few genes that
57 mitochondria and alphaproteobacteria still share. Second, the *Alphaproteobacteria* is under sampled and
58 most of its diversity remains to be discovered, as suggested by recent metagenomic surveys¹¹. Third, and
59 perhaps most problematic, the genomes of some lineages in the *Alphaproteobacteria* and those of
60 mitochondria have undergone convergent evolution. For example, the *Rickettsiales* and *Holosporaceae*
61 (intracellular bacteria)²⁹, or the *Pelagibacterales* and '*Puniceispirillaceae*' (planktonic bacteria)³⁰, have
62 reduced or streamlined genomes with compositionally biased genes similar to those of mitochondria. The
63 genes and genomes of these taxa are biased towards A+T nucleotides (and their proteins towards F, I, M,
64 N, K, and Y amino acids) in contrast to other groups that have not evolved reductively (which might be
65 biased towards G+C nucleotides and G, A, R, and P amino acids)²⁹. This sort of compositional
66 heterogeneity is often the cause of artefactual attractions among lineages with similar compositional
67 biases in phylogenetic inference³¹.

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

75 **Results**

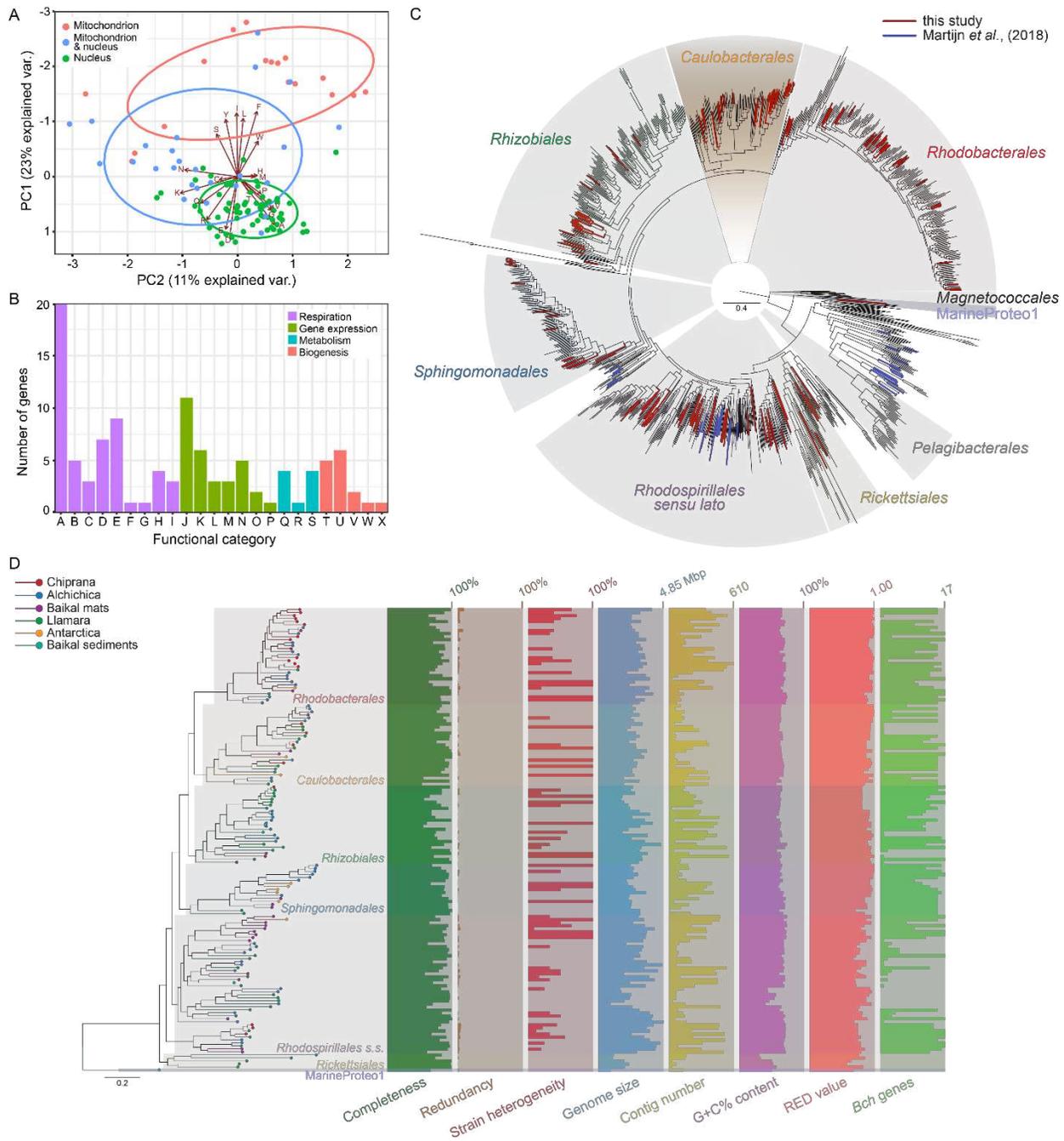
76 To date, most studies that aimed to phylogenetically place the mitochondrial lineage have relied
77 exclusively on mitochondrion-encoded protein datasets that range from 12 to 38 proteins^{16–18,32,11,12}.
78 These markers are not only few (e.g., 24 genes and 6,649 sites in ¹¹) but tend to be compositionally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

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

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

100 Most studies have exclusively relied on genomes of cultured alphaproteobacteria (e.g.,^{18–20,32}). Only one
101 recent study incorporated novel alphaproteobacterial MAGs from metagenomes sequenced by the Tara
102 Oceans project¹¹. So far, all of these alphaproteobacterial MAGs came from oceanic open waters and
103 tend to be small and A+T-rich¹¹. Moreover, none of them appeared to be most closely related to
104 mitochondria to the exclusion of other alphaproteobacteria¹¹.

105 To further increase taxonomic sampling across the *Alphaproteobacteria*, we assembled MAGs from
106 metagenomes sequenced from diverse microbial mats, microbialites, and lake sediments (see Table S2
107 for details). In addition, we also screened MAG collections released previously^{11,33–39}, as well as the
108 GTDB r89 database⁴⁰, for potentially phylogenetically novel alphaproteobacteria—together, these
109 databases comprise more than ~ 3,300 alphaproteobacterial genomes and MAGs. The newly assembled
110 MAGs were considerably diverse and widely distributed across the tree of the *Alphaproteobacteria* (Fig.
111 1C). Despite considerably expanding the sampled diversity of the *Alphaproteobacteria*, however, most of
112 these new MAGs appear to fall within previously sampled major clades (Fig. 1C, Fig. 1D, Table S3),
113 including those recently reported^{11,40} (Fig. 1D, Table S3). The most novel MAGs include new members of
114 the ‘early-diverging’ MarineProteo1 clade whose genomes are relatively small (1.43–2.71 Mbp) and not
115 heavily compositionally biased towards A+T (43.6–59.7%) (Fig. S2, Table S4). In addition, several novel
116 MAGs for ‘basal’ members of the *Rickettsiales* were found to be larger (1.47–2.36 Mbp) and enriched in
117 G+C (49.2–61.2 or ~49.3% on average) relative to previously sampled members of this group (>0.6–2.11
118 Mbp and 32.1–34.2% G+C on average in the *Rickettsiaceae*, *Anaplasmataceae*, and *Midichloriaceae*)
119 (Fig. S2, Table S4). The new alphaproteobacterial MAGs have moderate-to-high quality (according to
120 criteria by^{39,40}; 53.41–100% completeness and 0–9.17 redundancy), a wide range of G+C content (30.3–
121 73.5%) and sizes (0.88–4.85 Mbp), and varying degrees of phylogenetic novelty (0.99–0.56 Relative
122 Evolutionary Divergence score⁴⁰) (Fig. 1D, Table S3)—this suggests that the methods used here to
123 recover MAGs were not biased toward those with certain features (e.g., small sizes or high A+T content).
124 Most of the new MAGs, which are widely distributed across the *Alphaproteobacteria* tree, also appear to
125 encode an almost-complete set of bacteriochlorophyll biosynthesis enzymes which suggest that they
126 come from photosynthesizers in the diverse environments sampled (e.g., microbial mats; Fig. 1D, Table
127 S3).



128

129 **Figure 1. An expanded gene set and novel alphaproteobacterial MAGs from diverse environments.**

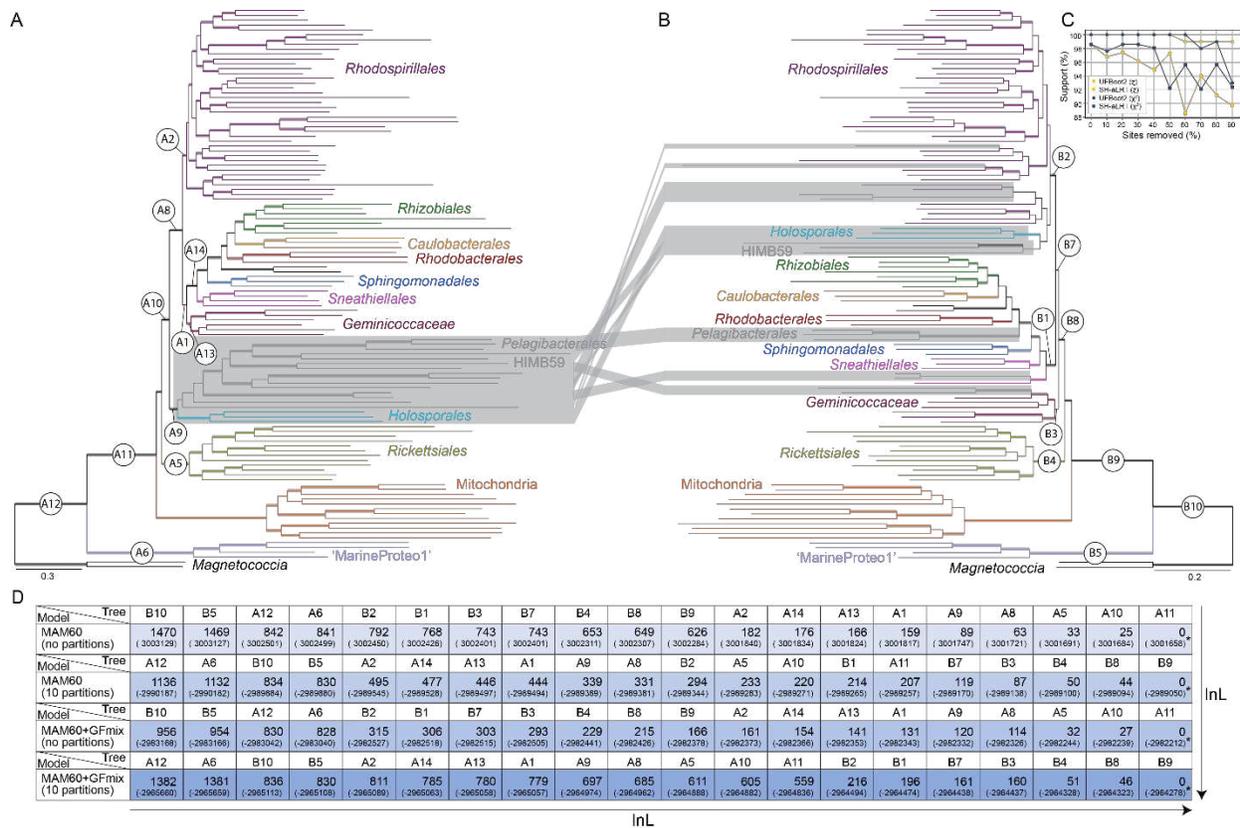
130 (A) Principal Component Analysis (PCA) of amino acid compositions for each one of the 108
 131 mitochondrial genes of alphaproteobacterial origin used in this study. Mitochondrion-encoded genes (light red); Mitochondrion- and nucleus-encoded genes (light blue); nucleus-encoded genes (green); 95%
 132 confidence ellipses follow the same color code as genes. This PCA was inferred from alignments that
 133 contain only eukaryotes. (B) Functional classification of the marker genes of alphaproteobacterial origin in
 134 eukaryotes used for multi-gene phylogenetic analyses in this study. All these functions take place inside
 135 mitochondria. A: Complex I subunit/assembly factor; B: Complex II subunit/assembly factor; C: Complex
 136 III subunit/assembly factor; D: Complex IV subunit/assembly factor; E: Complex V subunit/assembly
 137 factor; F: Cytochrome c biogenesis; G: D-lactate dehydrogenase (respiratory chain); H: Pyruvate
 138

139 dehydrogenase complex subunit; I: Krebs cycle; J: Ribosome large subunit; K: Ribosome small subunit;
140 L: Ribosome translational factor; M: rRNA modification/maturation; N: tRNA modification/maturation; O:
141 Aminoacyl-tRNA synthetase; P: RNA polymerase; Q: Branched-chain amino acid/fatty acid metabolism,
142 R: Pyrimidine biosynthesis; S: Ubiquinone biosynthesis; T: Protein import/export; U: Iron-sulfur cluster
143 biogenesis; V: Clp protease complex subunit; W: Proteasome-like complex subunit; X: Mitochondrial
144 division (see also Table S1). **(C)** Phylogenetic tree of 154 novel MAGs reported here, the 45 MAGs
145 reported by Martijn *et al.* (2018), and 1,188 of maximally diverse alphaproteobacterial genomes in GTDB
146 r89 database. Taxon sample reduction was done with Treemmer⁴¹ and phylogenetic inference with IQ-
147 TREE (-fast mode) and the LG4X model (120 GTDB-Tk marker genes; 14,048 amino acid sites). **(D)**.
148 Phylogenetic tree for the 154 alphaproteobacterial MAGs reconstructed from diverse metagenomes
149 sequenced in this study and summary of major features for each MAG. Tree was inferred with IQ-TREE (-
150 fast mode) and the LG4X model after having removed 50% of most compositionally heterogeneous ζ
151 sites (120 GTDB-Tk marker genes; 7,024 amino acid sites) (see also Table S3).

152 To address recent controversies^{11,26,12}, we first assembled the largest dataset to date that includes a new
153 set of 64 nucleus-encoded and 44 mitochondrion-encoded proteins (108 genes in total and 33,704 amino
154 acid sites; see above). Our dataset also comprised a wide taxon sampling with twelve mitochondria from
155 diverse eukaryotes (from most ‘supergroups’), and a broad set of 104 alphaproteobacteria that covered
156 all major known lineages and maximized phylogenetic diversity (subsampling from a set of more than
157 3,300 genomes; see Methods). Importantly, our dataset incorporated several *Rickettsiales* species that
158 have short branches and are less compositionally biased (Fig. 1D, Fig. S2, Table S4), as well as novel
159 representatives of the MarineProteo1 clade (Fig. 1D, Fig. 2A, Table S4). Instead of relying on *Beta-*, and
160 *Gammaproteobacteria* as outgroups (as in ^{11,12}), we used the much closer *Magnetococcia* which has
161 been consistently found to be sister to all other alphaproteobacteria (e.g., ^{11,12,20}). This was done to
162 decrease potential artefactual attractions between the long mitochondrial branch and distant outgroups, a
163 concern raised before^{11,26,12}. Furthermore, we also removed sites estimated to have undergone functional
164 divergence at the origin of mitochondria (these represented only 5.2% of all sites) using the FunDi mixture
165 model⁴². This was done to reduce potential artefacts from model misspecification as no phylogenetic
166 model currently available adequately captures such patterns of functional divergence in proteins.

167 We first analyzed our dataset using the MAM60 site-heterogeneous model that was specifically inferred
168 from our own dataset—this model has been shown to have a better fit than generic site-heterogeneous
169 models (e.g., C10-60)⁴³. Analyses on the untreated dataset (i.e., without compositionally heterogeneous
170 sites removed) placed mitochondria as sister to all of the *Alphaproteobacteria* with maximum support, i.e.,
171 both the monophyly of the *Alphaproteobacteria* and the *Alphaproteobacteria*-mitochondria clade were
172 fully supported (Fig. 2A). However, these analyses also recovered the grouping between the
173 *Pelagibacterales*, *Holosporaceae*, and other long-branching species (Fig. 2C, Mendeley Data) that, in
174 previous work²⁹, were shown to artefactually attract each other because of similar amino acid
175 compositional biases. A common strategy for dealing with compositional heterogeneity in the absence of
176 site-and-branch-heterogeneous models is to remove alignment sites based on metrics that quantify their
177 compositional heterogeneity^{11,12,29}. The progressive removal of the compositionally most heterogeneous
178 sites according to the ζ and χ^2 metrics^{11,29,44} disrupted compositional attractions and showed clear support
179 for the *Alphaproteobacteria*-sister hypothesis (Fig. 2B, Fig. 2C).

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



190

191 **Figure 2. Phylogenetic tree of the *Alphaproteobacteria* and mitochondria, and support from our**
 192 **new site-and-branch-heterogeneous model. (A)** Phylogenetic tree for the *Alphaproteobacteria*
 193 and mitochondria derived from a site-heterogeneous analyses of an untreated dataset. **(B)** Phylogenetic tree
 194 for the *Alphaproteobacteria* and mitochondria derived from a site-heterogeneous analysis of a dataset
 195 from which 50% of the most compositionally heterogeneous sites according to the ζ metric had been
 196 removed. The removal of this amount of ζ sites minimizes the variation of G A R P/F I M N K Y amino
 197 acid ratios across taxa (Table S5). The taxonomic labels follow the higher-level taxonomy outlined in ²⁹.
 198 Thickened branches represent branch support values of >90% SH-aLRT and >90% UFBboot2+NNI. **(C)**
 199 Variation in support values for the placement of mitochondria outside of the *Alphaproteobacteria* (SH-
 200 aLRT and UFBboot2+NNI) throughout the progressive removal of compositionally heterogenous sites
 201 according to the ζ and χ^2 metrics. Support for the branch that groups mitochondria with all
 202 alphaproteobacteria (but excludes MarineProteo1 and the *Magnetococcia*) is always maximum (i.e.,
 203 100% SH-aLRT /100% UFBboot2+NNI; Mendeley Data). **(D)** Heatmap table summarizing the differences
 204 in log-likelihoods (InL) relative to the highest log-likelihood for several alternative placements of
 205 mitochondria (A1-14 and B1-B12 in **(A)** and **(B)**; see Table S6 and Fig. S4 for all tree topologies) under a
 206 conventional site-heterogeneous model (MAM60) and our new site-and-branch-heterogeneous model
 207 (MAM60+GFmix). Models (rows) are arranged in increasing order (from top to bottom) according to InL
 208 values. For each model (row), tree topologies (columns) are arranged in increasing order (from left to
 209 right) according to InL values. Absolute log-likelihood values for each tree (A-T) under the different
 210 models tested are reported within parentheses. For all four models, all topologies other than the
 211 maximum-likelihood tree were rejected with p -values of < 0.0001 according to Bonferroni-corrected χ^2
 212 tests. See Table S6 for all tree topologies and datasets tested.

213 All studies to date have exclusively relied on either site-homogenous or purely site-heterogeneous
214 models (e.g., CAT in PhyloBayes or C60 in IQ-TREE)^{11,12,14–20,22,23,32}. Indeed, no tractable model that
215 accounts for compositional heterogeneity across branches and sites simultaneously is available; current
216 branch-heterogeneous models cannot be combined with site-heterogeneous models³¹, or are too
217 computationally intensive and suffer from convergence problems^{45,46}. To overcome these shortcomings,
218 we developed a model that captures the most important compositional heterogeneity in
219 alphaproteobacterial genomes— namely the variation in the G A R P/F I M N K Y amino acid ratio that is
220 driven by variation in G+C vs. A+T nucleotide content (see ²⁹). Our new branch-heterogeneous model,
221 GFmix, models the variation in the ratio of G A R P/F I M N K Y amino acid frequencies across the
222 phylogenetic tree in combination with conventional site-heterogeneous models (e.g., C10-60, MAM and
223 UDM models). Briefly, this model requires a rooted tree, and introduces a new parameter that represents
224 the G A R P/F I M N K Y ratio for every branch in a tree that is based on the amino acid compositions of
225 all taxa that descend from that branch (see Materials and Methods for details). These parameters, in turn,
226 adjust the frequencies of each site class in the site-profile mixture model resulting in a new transition rate
227 matrix, Q_c , for each mixture class for the given branch. We developed and implemented the new GFmix
228 model in a maximum likelihood framework.

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

247 The top three trees often favored by the MAM60+GFmix model (i.e., those with the highest likelihoods)
248 have mitochondria in adjacent branches: *Alphaproteobacteria*-sister (trees A11 and B9 in Fig. 2A and Fig.
249 2B), *Rickettsiales*-sister (trees A5 and B4 in Fig. 2A and Fig. 2B), and mitochondria as sister to all
250 alphaproteobacteria except the *Rickettsiales* (or *Caulobacteridae*-sister; trees A10 and B8 in Fig. 2A and
251 Fig. 2B)^{29,47}. However, Bonferroni-corrected χ^2 topology tests show that the optimal trees that display the
252 *Alphaproteobacteria*-sister relationship are significantly better than all trees with other positions for
253 mitochondria (see Fig. 2D). Even though the *Alphaproteobacteria*-sister relationship is also favored by the
254 MAM60+GFmix model for the mitochondrion-encoded protein dataset, the *Caulobacteridae*-sister
255 relationship cannot be rejected by the Bonferroni-corrected χ^2 tests (i.e., p -values > 0.05; Table S6). This
256 further supports the notion that the phylogenetic signal for the placement of mitochondria is weaker in
257 mitochondrion-encoded proteins (see above). The *Rickettsiales*-sister relationship is rejected for all
258 datasets and models (p -value < 0.005; Table S6). Overall, most of our distinct phylogenetic approaches
259 show support for the *Alphaproteobacteria*-sister hypothesis.

260 Discussion

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

277 Several previous studies have suggested that mitochondria were either sister to the *Rickettsiales*¹⁸⁻²⁰ or
278 phylogenetically embedded in a larger group comprised of both the *Rickettsiales* and the
279 *Holosporaceae*²⁰. These hypotheses implied that the mitochondrial ancestor may have been an
280 intracellular parasite: throughout its early evolution, the ancestor of mitochondria changed its function
281 from an energy parasite to an ATP-producing respiratory organelle¹⁸⁻²¹. The finding that mitochondria are
282 no longer phylogenetically associated to the *Rickettsiales* and are instead sister to the entire
283 *Alphaproteobacteria* clade makes a parasitic origin of mitochondria less plausible. However, the nature of
284 the mitochondrial ancestor remains poorly constrained. Future studies on species of the MarineProteo1
285 clade might shed some light on the early evolution of the *Alphaproteobacteria*, and possibly also on the
286 mitochondrial ancestor. However, we note that the MarineProteo1 clade is separated by a long branch
287 from the *Alphaproteobacteria* and mitochondria. Currently available genomes for the MarineProteo1 clade
288 are relatively small, but not necessarily compositionally biased, and suggest that these
289 alphaproteobacteria might be reduced and physiologically specialized (Fig. S2, Table S4).

290 Unravelling the deep evolutionary history of mitochondria is an inherently hard phylogenetic problem. One
291 of the main challenges is to properly account for the drastically different compositional biases across
292 anciently diversified lineages²⁹. Here, we have moved towards overcoming this major obstacle. Our newly
293 developed and implemented site-and-branch-heterogeneous model allowed us, for the first time, to test
294 different phylogenetic placements for mitochondria relative to the *Alphaproteobacteria* while accounting
295 for the drastic amino acid compositional changes that alphaproteobacterial and mitochondrial proteins
296 have undergone. A consistent view emerges from the combination of modelling and reducing
297 compositional heterogeneity: the *Alphaproteobacteria*-sister hypothesis is robust and unlikely to be
298 artefactual. However, we caution that the phylogenetic signal preserved in mitochondrion-encoded
299 proteins is weak and ambiguous. The recovery of the *Rickettsiales*-sister relationship in previous
300 studies^{11,12} may thus be result of ambiguous phylogenetic signal and long-branch attraction. Therefore,
301 we suggest that it is currently best to view mitochondria as an early offshoot of the alphaproteobacterial
302 lineage that diverged just prior to the diversification of known extant groups. This is suggested by the
303 short internal branch lengths between mitochondria and *Alphaproteobacteria* (see Fig. 2A, Fig. 2B) and is
304 supported by the shared presence of the Mitochondrial Contact Site and Cristae Organizing System (i.e.,
305 a Mitofilin domain-containing Mic60) in only mitochondria and the *Alphaproteobacteria*, but not in
306 members of the *Magnetococcia* and MarineProteo1^{48,49} (Fig. S2, Table S4). Future efforts should focus
307 on exploring diverse environments for unknown and extant alphaproteobacterial lineages that may be
308 more closely related to mitochondria.

309 **Materials and Methods**

310 Metagenomic sequencing and MAG assembly

311 Samples collected from (1) microbial mats in the Salada de Chiprana (Spain, December 2013), Salar de
312 Llamara⁵⁰, Lakes Bezymyanoe and Reid (Antarctica, January 2017) and several hot springs around
313 Lake Baikal (Southern Siberia, July 2017), (2) microbialites in Lake Alchichica⁵¹, and (3) sediments in
314 Lake Baikal, were fixed in ethanol (>70%) *in situ* and stored at -20°C as previously described⁵⁰. Total
315 DNA was purified from samples using the DNeasy PowerBiofilm Kit (QIAGEN, Germany) by following the
316 manufacturer's guidelines. DNA extracted from microbialite fragments was further cleaned using the
317 DNeasy PowerClean Cleanup Kit (QIAGEN, Germany) as previously described⁵². DNA was quantified
318 using Qubit®. DNA library preparation and sequencing were performed with an Illumina HiSeq2000 v3
319 (2x100 bp paired-end reads) by Beckman Coulter Genomics (Danvers, MA, USA), and with an Illumina
320 HiSeq2500 (2x125 bp paired-end reads) by Eurofins Genomics (Ebersberg, Germany). A summary of the
321 metagenomic libraries sequenced can be found in Table S2.

322 Raw Illumina short reads from all sequenced Illumina paired-end libraries were quality-assessed with
323 FastQC v.0.11.7 and quality-filtered with Trimmomatic v.0.36⁵³. Libraries made from samples from Lake
324 Alchichica and the Llamara saltern were processed with the following workflow. Libraries were individually
325 assembled, and technical replicates co-assembled (Table S2), with metaSPAdes v.3.10.0⁵⁴. Contigs
326 smaller than 2,500 bp in the (co-)assemblies were removed. Filtered reads were then individually mapped
327 onto each assembly with Bowtie2 to obtain contig coverages⁵⁵. Contigs were binned using MaxBin v.2.2.2
328 which relies on differential coverage across samples, tetranucleotide composition and single-copy marker
329 genes⁵⁶. The completeness and contamination of the bins reported by MaxBin v.2.2.2 were assessed with
330 CheckM v.1.0.12⁵⁷. Genome bins that were phylogenetically affiliated to the *Alphaproteobacteria* based
331 on the manual examination of the CheckM reference genome tree (itself based on the concatenation of
332 43 marker genes) were retained. Reads were then individually mapped onto each alphaproteobacterial
333 genome bin with Bowtie2. All paired and unpaired reads that successfully mapped to the
334 alphaproteobacterial bins were subsequently co-assembled with metaSPAdes. The resulting co-assembly
335 was processed through the Anvi'o metagenomic workflow⁵⁸. In brief, reads were mapped to the final
336 metaSPAdes co-assembly with Bowtie2 to obtain contig coverage values. DIAMOND searches⁵⁹ of
337 predicted proteins against the NCBI GenBank nr database were done to assign taxonomic affiliations to
338 each contig. CONCOCT⁶⁰, implemented in the Anvi'o suite, was used to bin the resulting metagenome.
339 Contigs were organized according to the composition and coverage by anvi-interactive. The predicted
340 CONCOCT2 bins were visualized and manually refined based on their composition, coverage, taxonomy
341 and completeness/redundancy. Libraries made from samples from Antarctica, the Chiprana saltern and
342 Lake Baikal were processed with the following workflow. Libraries from the same location or environment
343 type were co-assembled with MEGAHIT v.1.1.1⁶¹. Contigs smaller than 2,500 bp in the co-assemblies
344 were removed. Filtered reads were then individually mapped onto each co-assembly with Bowtie2 to
345 obtain contig coverages. Contigs were binned using three different binners (MetaBAT v.2.12.1⁶², MaxBin
346 2.2.4⁵⁶, CONCOCT⁶⁰) and their results were combined into consensus contigs bins with DAS Tool
347 v.1.1.0⁶³.

348 Marker protein selection

349 We built an expanded dataset of mitochondrion- and nucleus-encoded proteins of alphaproteobacterial
350 origin in eukaryotes. For the nucleus-encoded proteins, BLAST⁶⁴ similarity searches of all proteins
351 contained in the predicted proteomes of 13 representative eukaryotes were conducted against a
352 database of 176 prokaryotes (136 bacteria and 40 archaea). BLAST hits were clustered into homologous
353 families with a custom Perl script, aligned with MAFFT and the L-INS-I method⁶⁵, and then trimmed with
354 BMGE⁶⁶. Phylogenetic trees for each homologous gene family were inferred under the LG model in
355 RAxML v.8⁶⁷. These trees were then sorted based on the criterion that eukaryotes form a clade with
356 alphaproteobacteria. Manual inspection of the trees then followed to remove paralogs and contaminants.
357 For mitochondrion-encoded genes, mitochondrial clusters of orthologous genes (MitoCOGs)⁶⁸ that are
358 widespread among eukaryotes were used.

359 Both mitochondrion-, and nucleus-encoded candidate marker proteins were then compared through
360 BLAST searches against those reported previously by Wang and Wu (2015)²⁰ and Martijn *et al.*, (2018)¹¹.
361 Our dataset encompassed most proteins from these other datasets, with few exceptions. The non-
362 redundant and remaining candidate marker proteins comprising the union of these five datasets, were
363 then further screened phylogenetically. Using a representative eukaryotic (mitochondrial) query for each
364 marker gene, BLAST searches were done against a database that comprises 107 diverse bacteria
365 (representing 27 cultured phyla) and 23 diverse eukaryotes (representing 6 major groups); eukaryotes
366 were selected based on the availability of both mitochondrial and nuclear genomes or transcriptomes
367 (see Table S7). Homologues were aligned with MAFFT, alignments trimmed with TrimAl⁶⁹ and single-
368 protein trees inferred with IQ-TREE⁷⁰. The single-protein trees were inspected visually to remove
369 duplicates, paralogues, and any other visual outlier such as extremely divergent sequences. Single-
370 protein trees were then re-inferred from the curated alignments and visually inspected. Proteins for which
371 trees showed a sister relationship between eukaryotes and alphaproteobacteria were kept for further
372 analyses. Finally, these candidate marker proteins were annotated and further refined using the EggNOG
373 database and BLASTp searches. The final marker proteins set comprised 108 genes, 64 of which are
374 exclusively nucleus-encoded, 17 are exclusively mitochondrion-encoded, and 27 are both mitochondrion-
375 and nucleus-encoded (Fig. S1). The annotations confirm that all marker proteins are predicted to be
376 localized to mitochondria in eukaryotes (Table S1).

377 Dataset assembly

378 To increase taxon sampling as much as possible, MAGs reported in Anantharaman *et al.*, (2016)³³,
379 Graham *et al.*, (2018)³⁴, Delmont *et al.*, (2018)³⁵, Martijn *et al.*, (2018)¹¹, Mehrshad *et al.*, (2016)³⁶, Tully *et al.*,
380 *et al.*, (2017)³⁷, Tully *et al.*, (2018)³⁸ and Parks *et al.*, (2017)³⁹ were added to those reconstructed here (see
381 Metagenomic analyses). To improve the quality of our MAG selection, MAGs were analyzed with the
382 CheckM lineage workflow and those with quality values (completeness – 5x contamination) lower than 50
383 were discarded, just as done before by Parks *et al.*, (2017, 2018)^{39,40}. MAGs were then filtered according
384 to their taxonomic affiliation to the *Alphaproteobacteria*. A phylogenetic tree for all MAGs and all
385 *Proteobacteria* taxa in the GTDB r89 database⁴⁰ was inferred from 120 marker proteins, built-in in the
386 GTDB-Tk software, using IQ-TREE v.1.6.10⁷⁰ and the LG4X+F model. To increase phylogenetic
387 accuracy, a second tree was inferred with the LG+PMSF(C60)+G4+F using the LG4X tree as guide. All
388 MAGs that fell within the *Alphaproteobacteria* clade in the GTDB-Tk tree were chosen for subsequent
389 analyses. Together, these added up to more than 3,300 alphaproteobacteria. In order to reduce
390 computational burden, Treemmer v.0.1b was then used to reduce the number of alphaproteobacterial
391 taxa from the GTDB-TK tree while maximizing phylogenetic diversity⁴¹. The Treemmer analysis was
392 constrained so representatives from major clades, as visually identified, were retained. Finally, a set of
393 reference alphaproteobacteria (formally described species) were added, and long-branching
394 alphaproteobacteria were replaced by short-branching relatives.

395 To retrieve homologues, PSI-BLAST searches with either one, two, or three iterations using
396 representative mitochondrial (eukaryotic) query sequences for each marker protein were done against a
397 database that comprised all carefully selected predicted proteomes. To remove non-orthologous
398 sequences, homologous protein sets were retrieved for each marker protein, aligned with MAFFT,
399 trimmed with TrimAl and trees inferred with IQ-TREE. The single-protein trees were visually inspected to
400 remove duplicates, paralogues, and any other visual outlier such as extremely divergent sequences. The
401 curated homologous protein sets were finally aligned again with MAFFT v.7.3.10 and the L-INS-I method.
402 To increase phylogenetic signal by removing poorly aligned and non-homologous aligned regions, Divvier
403 v.1.0 was used with the -partial and -mincol options⁷¹. Only sites with more than 10% of data were
404 retained. To reduce incongruency among proteins due to, for example, lateral gene transfer, Phylo-MCOA
405 v.1.4⁷² was employed on single-protein trees with UFBoot2+NNI as branch support which were inferred
406 with IQ-TREE v.1.6.10 and the best-fitting model as identified by Model-Finder^{70,73}. Single-protein
407 alignments were concatenated with SequenceMatrix v.1.8⁷⁴.

408 Phylogenetic analyses using site-heterogeneous models

409 For multi-protein phylogenetic analyses on the supermatrix, trees were first inferred in IQ-TREE v.1.6.10
 410 under the LG4X+F model. The resulting site-homogenous tree was then used as a guide tree to infer a
 411 new phylogenetic tree under the LG+PMSF(C60)+F+G4 model⁷⁵. Consequently, the resulting site-
 412 heterogeneous tree was used as a guide tree to infer a new phylogenetic tree under the dataset-specific
 413 LG+PMSF(MAM60)+F+G4 model. The dataset-specific MAM60 model was estimated using the MAMMaL
 414 software⁴³. This site-heterogeneous mixture model is directly inferred from the dataset analyzed and
 415 therefore is more specific than the general C10-60 mixture models. To account for more than 60 (e.g.,
 416 C60 or MAM60) amino-acid composition profiles across the data, we used the general UDM128 mixture
 417 model as LG+UDM128+G4+F that allows for 128 amino acid composition profiles⁷⁶. The software FunDi
 418 was used to estimate functionally divergent sites in the branch that separates the mitochondrial lineage
 419 from all other taxa⁴². Sites with a probability of being functionally divergent > 0.5 were removed.
 420 Progressive removal of compositionally heterogeneous sites was performed according to the ζ and the χ^2
 421 metrics/methods as described before^{11,29,44}. Both metrics are designed to estimate compositional
 422 heterogeneity per site based on different criteria.

423 Bayesian analyses were conducted with PhyloBayes MPI v1.8 using the CAT-GTR+G4 model^{77,78}.
 424 PhyloBayes MCMC chains were run for >20,000 cycles or until convergence between the chains was
 425 achieved and the largest discrepancy in posterior probabilities for splits between chains ('max-diff') was
 426 <0.1. Individual chains were summarized into a Bayesian consensus tree using a burn-in of 500 trees and
 427 subsampling every 10 trees. However, most chains did not reach convergence or resolve the
 428 phylogenetic placement of mitochondria relative to alphaproteobacterial lineages (Mendeley Data).

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

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

433
$$P(\mathbf{x}_i; w_c, \theta) = \sum_{c=1}^C w_c \sum_{k=1}^K P(\mathbf{x}_i | r_k, \boldsymbol{\pi}^{(c)}; \theta) / K$$

434 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
 435 the site-profile mixture model, w_c is the class weight and θ is the vector of other adjustable parameters
 436 (branch lengths, α shape parameter and tree topology) in the model. In order to model shifts in the
 437 relative frequencies of the amino acids G A R P (specified by G+C-rich codons) and F I M N K Y
 438 (specified by A+T-rich codons) in different branches of the tree, the foregoing vectors of amino acid
 439 frequencies, $\boldsymbol{\pi}^{(c)}$, are modified in a branch-specific manner in the following way.

440 Let b denote the ratio of aggregate frequencies of G A R P to F I M N K Y amino acids; i.e., $b := \pi_G / \pi_F$
 441 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
 442 branch e in the phylogenetic tree under consideration, we can obtain estimates by a hierarchical
 443 procedure where b_e is obtained from the GARP/FIMNKY ratio of all of the sequences at the tips of the
 444 tree that descend from branch e . Using these estimates, the values in the class frequency vectors, $\boldsymbol{\pi}^{(c)}$,
 445 for any site profile class are modified in the following way to be branch- e -specific class frequencies, $\boldsymbol{\pi}^{(ce)}$.
 446 The modified class frequencies have to satisfy a number of constraints including:

447
$$\pi_j^{(ce)} = \begin{cases} \mu^{(ce)} S_G^{(e)} \pi_j^{(c)} & j \in \{G, A, R, P\} \\ \mu^{(ce)} S_F^{(e)} \pi_j^{(c)} & j \in \{F, Y, M, I, N, K\} \\ \mu^{(ce)} \pi_j^{(c)} & otherwise \end{cases}$$

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

449

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

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

455 Partitioning the data matrix for GFmix calculations.

456 The foregoing framework assumes that for each aligned protein in a given concatenated dataset, the
 457 GARP/FIMNKY ratios (b_e 's) for every branch in the tree will be similar. However, for our data matrix this
 458 assumption is not true as different proteins show different degrees of GARP/FIMNKY variation across
 459 taxa depending on the location of the corresponding gene (e.g., nucleus-encoded vs. mitochondrial-
 460 encoded) and degree of conservation. For this reason, we clustered the proteins in our dataset into
 461 groups in the following way. For each protein v and each taxon t we calculated the GARP/FIMNKY ratio,
 462 $b_v^{(t)} = \pi_G^{(t)} / \pi_F^{(t)}$. Then, we calculated the overall distance between these ratios for every pair of proteins u
 463 and v in the data matrix as $d_{u,v} = \sum_t |b_v^{(t)} - b_u^{(t)}| / N_{u,v}$ where $N_{u,v}$ is the total number of taxa for which
 464 sequences were available for both proteins (this normalization accounts for the differing amounts of
 465 missing data for different proteins). The proteins were then clustered based on $d_{u,v}$ distances using the
 466 UPGMA algorithm in MEGA-X⁷⁹ and 10 clusters were chosen as a computationally tractable number of
 467 partitions for further analysis. The GFmix model was then applied to these 10 partitions allowing for
 468 separate b_e values and branch lengths for each partition. The overall log-likelihoods for topologies were
 469 obtained as the sum of log-likelihoods of that topology over all partitions.

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

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

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

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

493 alternative hypothesis was selected based on the data rather than being fixed *a priori*, this can induce a
494 selection bias⁸¹. The Bonferroni approach uses a input set of trees and approximates the p -value when
495 $H_A: \tau = \hat{\tau}$ by the Bonferroni-corrected p -value one would obtain testing $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$, $i \in A$
496 where A is the set of input trees that are compatible with the consensus tree of τ_0 and $\hat{\tau}$.

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

506 Rephrasing the test as approximately the same as the result of multiple tests $H_0: \tau = \tau_0$ against $H_i: \tau = \tau_i$,
507 $i \in A$ lays bare that multiple testing is the source of selection bias. Bonferroni correction is a widely used
508 approach to adjusting for multiple testing. As one final approximation, rather than using the usual
509 Bonferroni-corrected p -value, $n(A) p(\hat{\tau})$, we use the exact correction had the p -values coming from the
510 tests been independent,

$$511 \quad 1 - [1 - p(\hat{\tau})]^{n(A)}.$$

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

515 Profile Hidden Markov Model (pHMM) searches

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

523 Data Availability

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

530 **Acknowledgements**

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

- 539 1. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria.
540 *Current Biology* **27**, R1177–R1192 (2017).
- 541 2. Stairs, C. W., Leger, M. M. & Roger, A. J. Diversity and origins of anaerobic metabolism in
542 mitochondria and related organelles. *Phil. Trans. R. Soc. B* **370**, 20140326 (2015).
- 543 3. Müller, M. *et al.* Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. *Microbiol.*
544 *Mol. Biol. Rev.* **76**, 444–495 (2012).
- 545 4. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).
- 546 5. Cavalier-Smith, T. Predation and eukaryote cell origins: A coevolutionary perspective. *The*
547 *International Journal of Biochemistry & Cell Biology* **41**, 307–322 (2009).
- 548 6. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*
549 **521**, 173–179 (2015).
- 550 7. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity.
551 *Nature* **541**, 353–358 (2017).
- 552 8. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of
553 eukaryotes. *Nature Reviews Microbiology* **15**, 711–723 (2017).
- 554 9. Gray, M. W. Mitochondrial Evolution. *Cold Spring Harb Perspect Biol* **4**, a011403 (2012).
- 555 10. Gray, M. W. Mosaic nature of the mitochondrial proteome: Implications for the origin and
556 evolution of mitochondria. *PNAS* **112**, 10133–10138 (2015).
- 557 11. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside
558 the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
- 559 12. Fan, L. *et al.* Phylogenetic analyses with systematic taxon sampling show that mitochondria
560 branch within Alphaproteobacteria. *Nature Ecology & Evolution* **4**, 1213–1219 (2020).
- 561 13. Viale, A. M. & Arakaki, A. K. The chaperone connection to the origins of the eukaryotic
562 organelles. *FEBS Letters* **341**, 146–151 (1994).
- 563 14. Andersson, S. G. E. *et al.* The genome sequence of *Rickettsia prowazekii* and the origin of
564 mitochondria. *Nature* **396**, 133–140 (1998).
- 565 15. Wu, M. *et al.* Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined
566 genome overrun by mobile genetic elements. *PLoS Biol.* **2**, E69 (2004).
- 567 16. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome Phylogenies Indicate a Meaningful
568 A-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales. *Mol*
569 *Biol Evol* **23**, 74–85 (2006).
- 570 17. Williams, K. P., Sobral, B. W. & Dickerman, A. W. A robust species tree for the
571 alphaproteobacteria. *J. Bacteriol.* **189**, 4578–4586 (2007).
- 572 18. Sasser, D. *et al.* Phylogenomic Evidence for the Presence of a Flagellum and *cbb3* Oxidase in
573 the Free-Living Mitochondrial Ancestor. *Mol Biol Evol* **28**, 3285–3296 (2011).
- 574 19. Wang, Z. & Wu, M. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy
575 parasite. *PLoS ONE* **9**, e110685 (2014).
- 576 20. Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of
577 mitochondria. *Sci Rep* **5**, 7949 (2015).
- 578 21. Ball, S. G., Bhattacharya, D. & Weber, A. P. M. Pathogen to powerhouse. *Science* **351**, 659–660
579 (2016).
- 580 22. Thrash, J. C. *et al.* Phylogenomic evidence for a common ancestor of mitochondria and the
581 SAR11 clade. *Scientific Reports* **1**, 13 (2011).
- 582 23. Georgiades, K., Madoui, M.-A., Le, P., Robert, C. & Raoult, D. Phylogenomic Analysis of
583 *Odyssella thessalonicensis* Fortifies the Common Origin of Rickettsiales, *Pelagibacter ubique* and
584 *Reclimonas americana* Mitochondrion. *PLoS ONE* **6**, e24857 (2011).
- 585 24. Abhishek, A., Bavishi, A., Bavishi, A. & Choudhary, M. Bacterial genome chimaerism and the
586 origin of mitochondria. *Can. J. Microbiol.* **57**, 49–61 (2011).
- 587 25. Thiery, T., Landan, G., Schenk, M., Dagan, T. & Martin, W. F. An Evolutionary Network of
588 Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial
589 Origin. *Genome Biol Evol* **4**, 466–485 (2012).
- 590 26. Gawryluk, R. M. R. Evolutionary Biology: A New Home for the Powerhouse? *Current Biology* **28**,
591 R798–R800 (2018).

- 592 27. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the Age of Eukaryotes: Evaluating
593 Evidence from Fossils and Molecular Clocks. *Cold Spring Harb Perspect Biol* **6**, a016139 (2014).
- 594 28. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and
595 eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
- 596 29. Muñoz-Gómez, S. A. *et al.* An updated phylogeny of the Alphaproteobacteria reveals that the
597 parasitic Rickettsiales and Holosporales have independent origins. *eLife* **8**, e42535 (2019).
- 598 30. Luo, H. Evolutionary origin of a streamlined marine bacterioplankton lineage. *ISME J* **9**, 1423–
599 1433 (2015).
- 600 31. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).
- 601 32. Rodríguez-Ezpeleta, N. & Embley, T. M. The SAR11 group of alpha-proteobacteria is not related
602 to the origin of mitochondria. *PLoS ONE* **7**, e30520 (2012).
- 603 33. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected
604 biogeochemical processes in an aquifer system. *Nature Communications* **7**, 13219 (2016).
- 605 34. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-
606 distributed bacterial phototroph. *The ISME Journal* **12**, 1861–1866 (2018).
- 607 35. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are
608 abundant in surface ocean metagenomes. *Nature Microbiology* **3**, 804–813 (2018).
- 609 36. Mehrshad, M., Amoozegar, M. A., Ghai, R., Shahzadeh Fazeli, S. A. & Rodriguez-Valera, F.
610 Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the Brackish Waters
611 of the Caspian Sea. *Appl Environ Microbiol* **82**, 1599–1612 (2016).
- 612 37. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled
613 genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**, e3558 (2017).
- 614 38. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-
615 assembled genomes from the global oceans. *Sci Data* **5**, (2018).
- 616 39. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially
617 expands the tree of life. *Nature Microbiology* **2**, 1533–1542 (2017).
- 618 40. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially
619 revises the tree of life. *Nature Biotechnology* **36**, 996–1004 (2018).
- 620 41. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of
621 diversity. *BMC Bioinformatics* **19**, 164 (2018).
- 622 42. Gaston, D., Susko, E. & Roger, A. J. A phylogenetic mixture model for the identification of
623 functionally divergent protein residues. *Bioinformatics* **27**, 2655–2663 (2011).
- 624 43. Susko, E., Lincker, L. & Roger, A. J. Accelerated Estimation of Frequency Classes in Site-
625 Heterogeneous Profile Mixture Models. *Molecular Biology and Evolution* **35**, 1266–1283 (2018).
- 626 44. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and
627 phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
- 628 45. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary
629 and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23**, 2058–2071 (2006).
- 630 46. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino acid replacement.
631 *Mol. Biol. Evol.* **25**, 842–858 (2008).
- 632 47. Ferla, M. P., Thrash, J. C., Giovannoni, S. J. & Patrick, W. M. New rRNA gene-based
633 phylogenies of the Alphaproteobacteria provide perspective on major groups, mitochondrial ancestry
634 and phylogenetic instability. *PLoS ONE* **8**, e83383 (2013).
- 635 48. Muñoz-Gómez, S. A. *et al.* Ancient Homology of the Mitochondrial Contact Site and Cristae
636 Organizing System Points to an Endosymbiotic Origin of Mitochondrial Cristae. *Current Biology* **25**,
637 1489–1495 (2015).
- 638 49. Muñoz-Gómez, S. A., Wideman, J. G., Roger, A. J. & Slamovits, C. H. The Origin of
639 Mitochondrial Cristae from Alphaproteobacteria. *Mol. Biol. Evol.* **34**, 943–956 (2017).
- 640 50. Gutiérrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic
641 transitions. *Nature Ecology & Evolution* **2**, 1700–1708 (2018).
- 642 51. Saghaï, A. *et al.* Comparative metagenomics unveils functions and genome features of
643 microbialite-associated communities along a depth gradient. *Environ Microbiol* **18**, 4990–5004 (2016).
- 644 52. Saghaï, A. *et al.* Metagenome-based diversity analyses suggest a significant contribution of non-
645 cyanobacterial lineages to carbonate precipitation in modern microbialites. *Front Microbiol* **6**, 797
(2015).

- 647 53. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.
648 *Bioinformatics* **30**, 2114–2120 (2014).
- 649 54. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell
650 sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
- 651 55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**,
652 357–359 (2012).
- 653 56. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to
654 recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
- 655 57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing
656 the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome*
657 *Res.* **25**, 1043–1055 (2015).
- 658 58. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*
659 **3**, e1319 (2015).
- 660 59. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND.
661 *Nature Methods* **12**, 59–60 (2015).
- 662 60. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods*
663 **11**, 1144–1146 (2014).
- 664 61. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node
665 solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*
666 **31**, 1674–1676 (2015).
- 667 62. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
668 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
- 669 63. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation
670 and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).
- 671 64. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database
672 search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997).
- 673 65. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple
674 sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
- 675 66. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software
676 for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*
677 **10**, 210 (2010).
- 678 67. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
679 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
- 680 68. Kannan, S., Rogozin, I. B. & Koonin, E. V. MitoCOGs: clusters of orthologous genes from
681 mitochondria and implications for the evolution of eukaryotes. *BMC Evolutionary Biology* **14**, 237
682 (2014).
- 683 69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment
684 trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
- 685 70. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective
686 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274
687 (2015).
- 688 71. Ali, R. H., Bogusz, M. & Whelan, S. Identifying Clusters of High Confidence Homologies in
689 Multiple Sequence Alignments. *Mol Biol Evol* **36**, 2340–2351 (2019).
- 690 72. de Vienne, D. M., Ollier, S. & Aguilera, G. Phylo-MCOA: a fast and efficient method to detect
691 outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol. Biol. Evol.* **29**,
692 1587–1598 (2012).
- 693 73. Kalyanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder:
694 fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).
- 695 74. Vaidya, G., Lohman, D. J. & Meier, R. SequenceMatrix: concatenation software for the fast
696 assembly of multi-gene datasets with character set and codon information. *Cladistics* **27**, 171–180
697 (2011).
- 698 75. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior
699 Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol* **67**, 216–235
700 (2018).
- 701 76. Schrempf, D., Lartillot, N. & Szöllősi, G. Scalable Empirical Mixture Models That Account for
702 Across-Site Compositional Heterogeneity. *Molecular Biology and Evolution* **37**, 3616–3631 (2020).

- 703 77. Lartillot, N. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid
704 Replacement Process. *Molecular Biology and Evolution* **21**, 1095–1109 (2004).
- 705 78. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction
706 with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
- 707 79. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics
708 Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547–1549 (2018).
- 709 80. Susko, E. Tests for Two Trees Using Likelihood Methods. *Molecular Biology and Evolution* **31**,
710 1029–1039 (2014).
- 711 81. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to
712 Phylogenetic Inference. *Molecular Biology and Evolution* **16**, 1114–1114 (1999).
- 713 82. Markowski, E. A comparison of methods for constructing confidence sets of phylogenetic trees
714 using maximum likelihood. (Dalhousie University, 2021).
- 715

Supplementary Files

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

- [TableS1.xlsx](#)
- [TableS2.xlsx](#)
- [TableS3.xlsx](#)
- [TableS4.xlsx](#)
- [TableS5.xlsx](#)
- [TableS6.xlsx](#)
- [TableS7.xlsx](#)
- [SupplementaryMaterial20210523.docx](#)