

 Open access • Posted Content • DOI:10.1101/2021.01.19.427276

An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes — Source link

Edmund R. R. Moody, Tara A. Mahendrarajah, Nina Dombrowski, James W. Clark ...+5 more authors

Institutions: University of Bristol, Eötvös Loránd University, Science for Life Laboratory

Published on: 07 May 2021 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Phylogenetic tree and Ribosomal protein

Related papers:

- [Universal markers support a long inter-domain branch between Archaea and Bacteria](#)
- [Archaeal Phylogeny Based on Ribosomal Proteins](#)
- [Genome trees constructed using five different approaches suggest new major bacterial clades](#)
- [Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes.](#)
- [A congruent phylogenomic signal places eukaryotes within the Archaea](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/an-estimate-of-the-deepest-branches-of-the-tree-of-life-from-2uie0ct3o2>

1 An estimate of the deepest branches of the tree of 2 life from ancient vertically-evolving genes

3
4 Edmund R. R. Moody¹, Tara A. Mahendrarajah², Nina Dombrowski², James W. Clark¹, Celine
5 Petitjean¹, Pierre Offre², Gergely J. Szöllősi^{3,4,5}, Anja Spang^{2,6*}, Tom A. Williams^{1*}

- 6
7 1. School of Biological Sciences, University of Bristol, Bristol BS8 1TH, UK.
8 2. NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine
9 Microbiology and Biogeochemistry; AB Den Burg, The Netherlands
10 3. Dept. of Biological Physics, Eötvös Loránd University, 1117 Budapest, Hungary
11 4. MTA-ELTE “Lendület” Evolutionary Genomics Research Group, 1117 Budapest,
12 Hungary;
13 5. Institute of Evolution, Centre for Ecological Research, 1121 Budapest, Hungary
14 6. Department of Cell- and Molecular Biology, Science for Life Laboratory, Uppsala
15 University, SE-75123, Uppsala, Sweden

16
17 *Co-corresponding authors: tom.a.williams@bristol.ac.uk, anja.spang@nioz.nl

18 Abstract

19
20 Core gene phylogenies provide a window into early evolution, but different gene sets and
21 analytical methods have yielded substantially different views of the tree of life. Trees inferred
22 from a small set of universal core genes have typically supported a long branch separating
23 the archaeal and bacterial domains. By contrast, recent analyses of a broader set of non-
24 ribosomal genes have suggested that Archaea may not be very divergent from Bacteria, and
25 that estimates of inter-domain distance are inflated due to accelerated evolution of ribosomal
26 proteins along the inter-domain branch. Resolving this debate is key to determining the
27 diversity of the archaeal and bacterial domains, the shape of the tree of life, and our
28 understanding of the early course of cellular evolution. Here, we investigate the evolutionary
29 history of the marker genes key to the debate. We show that estimates of a reduced Archaea-
30 Bacteria (AB) branch length result from inter-domain gene transfers and hidden paralogy in
31 the expanded marker gene set, which act to artifactually diminish the genetic distance between
32 the two domains. By contrast, analysis of a broad range of manually curated marker gene
33 datasets from a sample of 700 Archaea and Bacteria reveal that current methods likely
34 underestimate the AB branch length due to substitutional saturation and poor model fit; that
35 the best-performing phylogenetic markers tend to support longer inter-domain branch lengths;
36 and that the AB branch lengths of ribosomal and non-ribosomal marker genes are statistically
37 indistinguishable. A phylogeny of prokaryotes inferred from the 27 highest-ranked marker
38 genes, including ribosomal and non-ribosomal markers, supported a long AB branch,
39 recovered a clade of DPANN at the base of the Archaea, and placed CPR within Bacteria as
40 the sister group to the Chloroflexota.

41 Introduction

42

43 Much remains unknown about the earliest period of cellular evolution and the deepest
44 divergences in the tree of life. Phylogenies encompassing both Archaea and Bacteria have
45 been inferred from a “universal core” set of 16-56 genes encoding proteins involved in
46 translation and other aspects of the genetic information processing machinery (Ciccarelli et al.,
47 2006; Fournier and Gogarten, 2010; Harris et al., 2003; Hug et al., 2016; Mukherjee et al.,
48 2017; Petitjean et al., 2014; Ramulu et al., 2014; Raymann et al., 2015; Theobald, 2010;
49 Williams et al., 2020). These genes are thought to predominantly evolve vertically and are thus
50 best-suited for reconstructing the tree of life (Ciccarelli et al., 2006; Creevey et al., 2011;
51 Ramulu et al., 2014; Theobald, 2010). In these analyses, the branch separating Archaea from
52 Bacteria (hereafter, the AB branch) is often the longest internal branch in the tree (Cox et al.,
53 2008; Gogarten et al., 1989; Hug et al., 2016; Iwabe et al., 1989; Pühler et al., 1989; Williams
54 et al., 2020). In molecular phylogenetics, branch lengths are usually measured in expected
55 numbers of substitutions per site, with a long branch corresponding to a greater degree of
56 genetic change. Long branches can therefore result from high evolutionary rates, long periods
57 of absolute time, or a combination of the two. If a sufficient number of fossils are available to
58 calibrate them then molecular clock models can, in principle, disentangle the contributions of
59 these effects. However, only very few fossil calibrations (Sugitani et al., 2015) are currently
60 available that are old enough to calibrate early divergences (Betts et al., 2018; Horita and
61 Berndt, 1999; Lepland et al., 2002; van Zuilen et al., 2002), and as a result, the ages and
62 evolutionary rates of the deepest branches of the tree, and estimates of the true biodiversity
63 of the archaeal and bacterial domains, remain highly uncertain.

64

65 Recently, Zhu et al. (Zhu et al., 2019) inferred a phylogeny from 381 genes distributed across
66 Archaea and Bacteria using the supertree method ASTRAL (Mirarab et al., 2014). In addition
67 to a large increase in the number of genes compared to other universal marker sets, the
68 functional profile of these markers comprises not only proteins involved in information
69 processing but also proteins affiliated with most other functional COG categories, including
70 metabolic processes (Table S1). The genetic distance (branch length) between the domains
71 (Zhu et al., 2019) was estimated from a concatenation of the same marker genes, resulting in
72 a much shorter AB branch length than observed with the core universal markers (Hug et al.,
73 2016; Williams et al., 2020). These analyses were consistent with the hypothesis (Petitjean et
74 al., 2014; Zhu et al., 2019) that the apparent deep divergence of Archaea and Bacteria might
75 be the result of an accelerated evolutionary rate of genes encoding translational and in
76 particular ribosomal proteins along the AB branch as compared to other genes. Interestingly,
77 the same observation was made previously using a smaller set of 38 non-ribosomal marker
78 proteins (Petitjean et al., 2014), although the difference in AB branch length between
79 ribosomal and non-ribosomal markers in that analysis was reported to be substantially lower
80 (roughly two-fold, compared to roughly ten-fold for the 381 protein set (Petitjean et al., 2014;
81 Zhu et al., 2019).

82

83 A higher evolutionary rate of ribosomal genes might result from the accumulation of
84 compensatory substitutions at the interaction surfaces among the protein subunits of the
85 ribosome (Petitjean et al., 2014; Valas and Bourne, 2011), or as a compensatory response to
86 the addition or removal of ribosomal subunits early in evolution (Petitjean et al., 2014).

87 Alternatively, differences in the inferred AB branch length might result from varying rates or
88 patterns of evolution between the traditional core genes (Spang et al., 2015; Williams et al.,
89 2020) and the expanded set (Zhu et al., 2019). Substitutional saturation (multiple substitutions
90 at the same site (Jeffroy et al., 2006)) and across-site compositional heterogeneity can both
91 impact the inference of tree topologies and branch lengths (Foster, 2004; Lartillot et al., 2007;
92 Lartillot and Philippe, 2004; Quang et al., 2008; Wang et al., 2008). These difficulties are
93 particularly significant for ancient divergences (Gouy et al., 2015). Failure to model site-
94 specific amino acid preferences has previously been shown to lead to under-estimation of the
95 AB branch length due to a failure to detect convergent changes (Tourasse and Gouy, 1999;
96 Williams et al., 2020), although the published analysis of the 381 marker set did not find
97 evidence of a substantial impact of these features on the tree as a whole (Zhu et al., 2019).
98 Those analyses also identified phylogenetic incongruence among the 381 markers, but did
99 not determine the underlying cause (Zhu et al., 2019).

100

101 This recent work (Zhu et al., 2019) raises two important issues regarding the inference of the
102 universal tree: first, that estimates of the genetic distance between Archaea and Bacteria from
103 classic “core genes” may not be representative of ancient genomes as a whole, and second,
104 that there may be many more suitable genes to investigate early evolutionary history than
105 generally recognized, providing an opportunity to improve the precision and accuracy of deep
106 phylogenies. Here, we investigate these issues in order to determine why different marker sets
107 support different Archaea-Bacteria branch lengths. First, we examine the evolutionary history
108 of the 381 gene marker set (hereafter, the expanded marker gene set) and identify several
109 features of these genes, including instances of inter-domain gene transfers and mixed
110 paralogy, that may contribute to the inference of a shorter AB branch length in supertree and
111 concatenation analyses. Then, we re-evaluate the marker gene sets used in a range of
112 previous analyses to determine how these and other factors, including substitutional saturation
113 and model fit, contribute to inter-domain branch length estimations and the shape of the
114 universal tree. Finally, we identify a subset of marker genes least affected by these issues,
115 and use these to estimate an updated tree of the primary domains of life and the genetic
116 distance between Archaea and Bacteria.

117 Results and Discussion

118 Gene transfers and hidden paralogy obscure the genetic 119 distance between Archaea and Bacteria

120

121 ***Genes from the expanded marker set are not widely distributed in Archaea***

122

123 The 381 gene set was derived from a larger set of 400 genes used to estimate the phylogenetic
124 placement of new lineages as part of the PhyloPhlAn method (Segata et al., 2013). Perhaps
125 reflecting the focus on bacteria in the original application, the phylogenetic distribution of the
126 381 marker genes in the expanded set varies substantially (Table S1), with many being poorly
127 represented in Archaea. Indeed 25% of the published gene trees (<https://biocore.github.io/wol/>
128 (Zhu et al., 2019)) contain less than 0.5% archaeal homologues, with 21 (5%) and 69 (18%)
129 of these trees including no or less than 10 archaeal homologues, respectively. For the
130 remaining 75% of the gene trees, archaeal homologs comprise 0.5%-13.4% of the dataset.
131 While there are many more sequenced bacteria than archaea, 63% of the gene trees
132 possessed genes from less than half of the 669 archaeal genomes included in the analysis,
133 whereas only 22% of the gene trees possessed fewer than half of the total number of 9906
134 sampled bacterial genomes. These distributions suggest that many of these genes are not
135 broadly present in both domains, and that some might be specific to Bacteria.

136

137

138 ***Conflicting evolutionary histories of individual marker genes and the inferred species 139 tree***

140

141 In the focal analysis of the 381 gene set, the tree topology was inferred using the supertree
142 method ASTRAL (Mirarab et al., 2014), with branch lengths inferred on this fixed tree from a
143 marker gene concatenation (Zhu et al., 2019). The topology inferred from this expanded
144 marker set (Zhu et al., 2019) is similar to published trees (Castelle and Banfield, 2018; Hug et
145 al., 2016) and recovers Archaea and Bacteria as reciprocally monophyletic domains, albeit
146 with a shorter AB branch than in earlier analyses. However, the individual gene trees (Zhu et
147 al., 2019) disagree regarding domain monophyly: Archaea and Bacteria are recovered as
148 reciprocally monophyletic groups in only 24 of the 381 published (Zhu et al., 2019) maximum
149 likelihood (ML) gene trees of the expanded marker set (Table S1).

150

151 Since single gene trees often fail to strongly resolve ancient relationships, we used
152 approximately-unbiased (AU) tests (Shimodaira, 2002) to evaluate whether the failure to
153 recover domain monophyly in the published ML trees is statistically supported. For
154 computational tractability, we performed these analyses on a 1000-species subsample of the
155 full 10,575-species dataset that was compiled in the original study (Zhu et al., 2019). For 79
156 of the 381 genes, we could not perform the test because the gene was not found on any of
157 the 74 archaeal genomes present in the 1000-species subsample. For the remaining 302
158 genes, domain monophyly was rejected ($p < 0.05$) for 232 out of 302 (76.8%) genes. As a
159 comparison, we performed the same test on several smaller marker sets used previously to
160 infer a tree of life (Coleman et al., 2021; Petitjean et al., 2014; Williams et al., 2020); none of

161 the markers in those sets rejected reciprocal domain monophyly ($p > 0.05$ for all genes, Figure
162 1(a)). In what follows, we refer to four published marker gene sets as: the Expanded set (381
163 genes (Zhu et al., 2019)), the Core set (49 genes (Williams et al., 2020), encoding ribosomal
164 proteins and other conserved information-processing functions; itself a consensus set of
165 several earlier studies (Da Cunha et al., 2017; Spang et al., 2015; Williams et al., 2012)), the
166 Non-ribosomal set (38 genes, broadly distributed and explicitly selected to avoid genes
167 encoding ribosomal proteins (Petitjean et al., 2014)), and the Bacterial set (29 genes used in
168 a recent analysis of bacterial phylogeny (Coleman et al., 2021)).

169
170 To investigate why 232 of the marker genes rejected the reciprocal monophyly of Archaea and
171 Bacteria, we returned to the full dataset (Zhu et al., 2019), annotated each sequence in each
172 marker gene family by assigning proteins to KOs, Pfams, and Interpro domains, among others
173 (Table S1, see Methods for details) and manually inspected the tree topologies (Table S1).
174 This revealed that the major cause of domain polyphyly observed in gene trees was inter-
175 domain gene transfer (in 357 out of 381 gene trees (93.7%)) and mixing of sequences from
176 distinct paralogous families (in 246 out of 381 gene trees (64.6%)). For instance, marker
177 genes encoding ABC-type transporters (p0131, p0151, p0159, p0174, p0181, p0287, p0306,
178 p0364), tRNA synthetases (i.e. p0000, p0011, p0020, p0091, p0094, p0202),
179 aminotransferases and dehydratases (i.e. p0073/4-aminobutyrate aminotransferase;
180 p0093/3-isopropylmalate dehydratase) often comprised a mixture of paralogues.

181
182 Together, these analyses indicate that the evolutionary histories of the individual markers of
183 the expanded set differ from each other and from the species tree. Zhu et al. acknowledged
184 (Zhu et al., 2019) the varying levels of congruence between the marker phylogenies and the
185 species tree, but did not investigate the underlying causes. Our analyses establish the basis
186 for these disagreements in terms of gene transfers and the mixing of orthologues and
187 paralogues within and between domains. Concatenation is based on the assumption that all
188 of the genes in the supermatrix evolve on the same underlying tree; genes with different gene
189 tree topologies violate this assumption and should not be concatenated because the
190 topological differences among sites are not modelled, and so the impact on inferred branch
191 lengths is difficult to predict. In practice, it is often difficult to be certain that all of the markers
192 in a concatenate share the same gene tree topology, and the analysis proceeds on the
193 hypothesis that a small proportion of discordant genes are not expected to seriously impact
194 the inferred tree. However, the concatenated tree inferred from the expanded marker set
195 differs from previous trees in that the genetic distance between Bacteria and Archaea is greatly
196 reduced, such that the AB branch length appears comparable to distances among bacterial
197 phyla (Zhu et al., 2019). Because an accurate estimate of the AB branch length has a major
198 bearing on unanswered questions regarding the root of the universal tree (Gouy et al., 2015),
199 we next evaluated the impact of the conflicting gene histories within the expanded marker set
200 on inferred AB branch length.

201
202 ***The inferred branch length between Archaea and Bacteria is artifactually shortened by***
203 ***inter-domain gene transfer and hidden paralogy***

204
205 To investigate the impact of gene transfers and mixed paralogy on the AB branch length
206 inferred by gene concatenations (Zhu et al., 2019), we compared branch lengths estimated
207 from markers that rejected ($AU < 0.05$) or did not reject ($AU > 0.05$) the reciprocal monophyly
208 of Bacteria and Archaea in the 381 marker set (Figure 1(a)). To estimate AB branch lengths

209 for genes in which the domains were not monophyletic in the ML tree, we first performed a
210 constrained ML search to find the best gene tree that was consistent with domain monophyly
211 for each family under the LG+G4+F model in IQ-TREE 2 (Minh et al., 2020). While it may
212 seem strained to estimate the length of a branch that does not appear in the ML tree, we
213 reasoned that this approach would provide insight into the contribution of these genes to the
214 AB branch length in the concatenation, in which they conflict with the overall topology. AB
215 branch lengths were significantly ($P = 2.159 \times 10^{-12}$, Wilcoxon rank sum test) shorter for markers
216 that rejected domain monophyly (Figure 1(a); <0.05 : mean AB branch length in expected
217 substitutions/site 0.0130, >0.05 : mean AB branch length 0.559). This result suggests that
218 inter-domain gene transfers reduce the AB branch length when included in a concatenation.
219 This behaviour might result from marker gene transfers reducing the number of fixed
220 differences between the domains, so that the AB branch length in a tree in which Archaea and
221 Bacteria are constrained to be reciprocally monophyletic will tend towards 0 as the number of
222 transfers increases. Consistent with this hypothesis, we observed that ΔLL , the difference in
223 log likelihood between the constrained ML tree and ML gene tree (used here as a proxy for
224 gene verticality), correlates negatively with AB branch length (Figure 1(b)). Furthermore, AB
225 branch length decreased as increasing numbers of low-verticality markers were added to the
226 concatenate (Figure 1(c)). Taken together, these results indicate that the inclusion of genes
227 that do not support the reciprocal monophyly of Archaea and Bacteria in the universal
228 concatenate reduces the estimated AB branch length by homogenizing the genetic diversity
229 of the two domains.

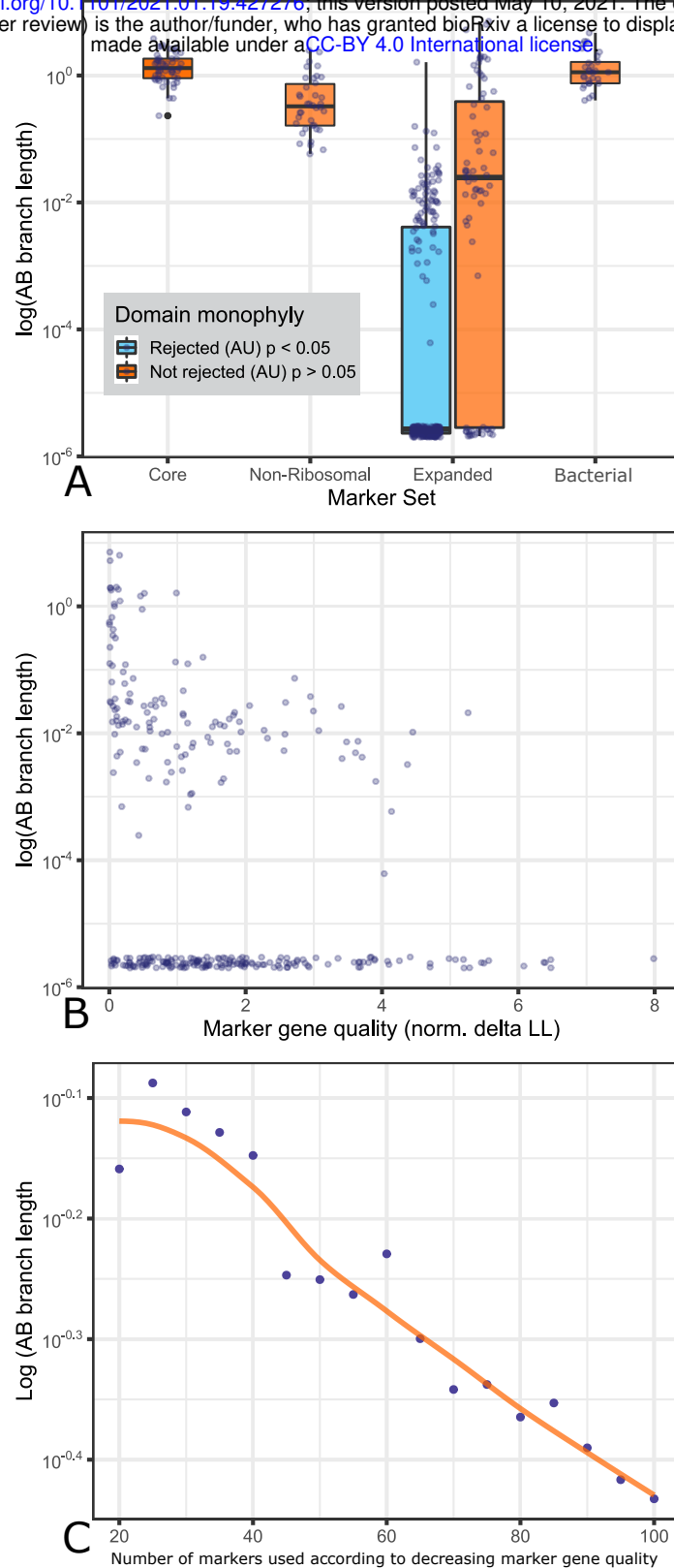


Figure 1: Expanded set genes in which Archaea and Bacteria are not monophyletic support a shorter AB branch. (a) Expanded set genes that reject domain monophyly ($p < 0.05$, AU test) support significantly shorter AB branch lengths when constrained to follow a domain monophyletic tree ($p = 2.159 \times 10^{-12}$, Wilcoxon rank-sum test). None of the marker genes from several other published analyses reject domain monophyly ($p > 0.05$, AU test) for all genes tested. (b) Marker gene verticality (ΔLL , see below) for the expanded gene set normalized by alignment length correlates negatively with the length of the AB branch between Archaea and Bacteria ($R^2 = 0.03998$, $p = 0.0004731$). (c) Concatenations of 20-100 markers of the expanded set markers ranked by marker gene verticality (ΔLL) show the same trend, with a reduction in AB branch length as markers with a greater ΔLL are added to the concatenate. ΔLL is the difference between the log likelihood of the ML gene family tree under a free topology search and the log likelihood of the best tree constrained to obey domain monophyly. The trendline is estimated using LOESS regression.

230 ***The age of the last universal common ancestor (LUCA) inferred from strict clocks***
231 ***does not predict marker gene quality***

232

233 Reliable age estimates using molecular clock methods require calibrations, but few
234 calibrations exist for the deeper branches of the tree of life. Zhu et al. (Zhu et al., 2019) argued
235 that the expanded marker set is useful for deep phylogeny because estimates of the age of
236 the last universal common ancestor (LUCA) obtained by fitting molecular clocks to their
237 dataset are in agreement with the geological record: a root (LUCA) age of 3.6-4.2 Ga was
238 inferred from the entire 381-gene dataset, consistent with the earliest fossil evidence for life
239 (Betts et al., 2018; Sugitani et al., 2015), whereas estimates from ribosomal markers alone
240 supported a root age of 7 Ga. This age might be considered implausible because it is much
241 older than the age of the Earth and Solar System (with the moon-forming impact occurring
242 ~4.51 Ga (Barboni et al., 2017; Hanan and Tilton, 1987)). However, the palaeobiological
243 plausibility of the age estimate from the 381 gene set does not, in itself, constitute evidence of
244 marker gene suitability. In the original analyses, the age of LUCA was estimated using a
245 maximum likelihood approach, as well as a Bayesian molecular clock with a strict clock
246 (assuming a constant evolutionary rate) or a relaxed clock with a single calibration. A strict
247 clock model does not permit changes in evolutionary rate through time or across branches,
248 and so a longer AB branch will lead to an older inferred LUCA age. Likewise, a relaxed clock
249 model with a single calibration may fail to distinguish molecular distances and geological time.
250 Given that the short AB branch in the expanded gene set results, in part, from phylogenetic
251 incongruence among markers, we evaluated the age of LUCA inferred from the subset of the
252 expanded gene set least affected by these issues. To do so, we analysed the top 5% of gene
253 families according to their ΔLL score (a set of 20 genes, which includes only 1 ribosomal
254 protein) under the same clock model parameters as the original dataset (Figure 2). This
255 analysis resulted in a significantly more ancient age estimate for LUCA (5.5-6.5 Ga), and
256 trimming the alignment to remove poorly-aligning regions resulted in a still older estimate
257 (6.34-6.89 Ga), approaching that of the ribosomal genes (7.46-8.03 Ga). These analyses
258 suggest that, for these data and calibrations, the inferred age for LUCA is not a reliable
259 indicator of marker quality, because analyses using the subset of the data least affected by
260 incongruence more clearly reveals the underlying limitations of strict clock analyses (and
261 indeed relaxed clocks with few calibrations) for dating ancient divergences. In principle, more
262 reliable estimates of LUCA's age might be obtained by using more calibrations. However,
263 unambiguous calibrations remain elusive, particularly for the root and other deep branches of
264 the tree. Despite advances in molecular clock methodology, such calibrations represent the
265 only way to reliably capture the relationship between genetic distance and divergence time.

266

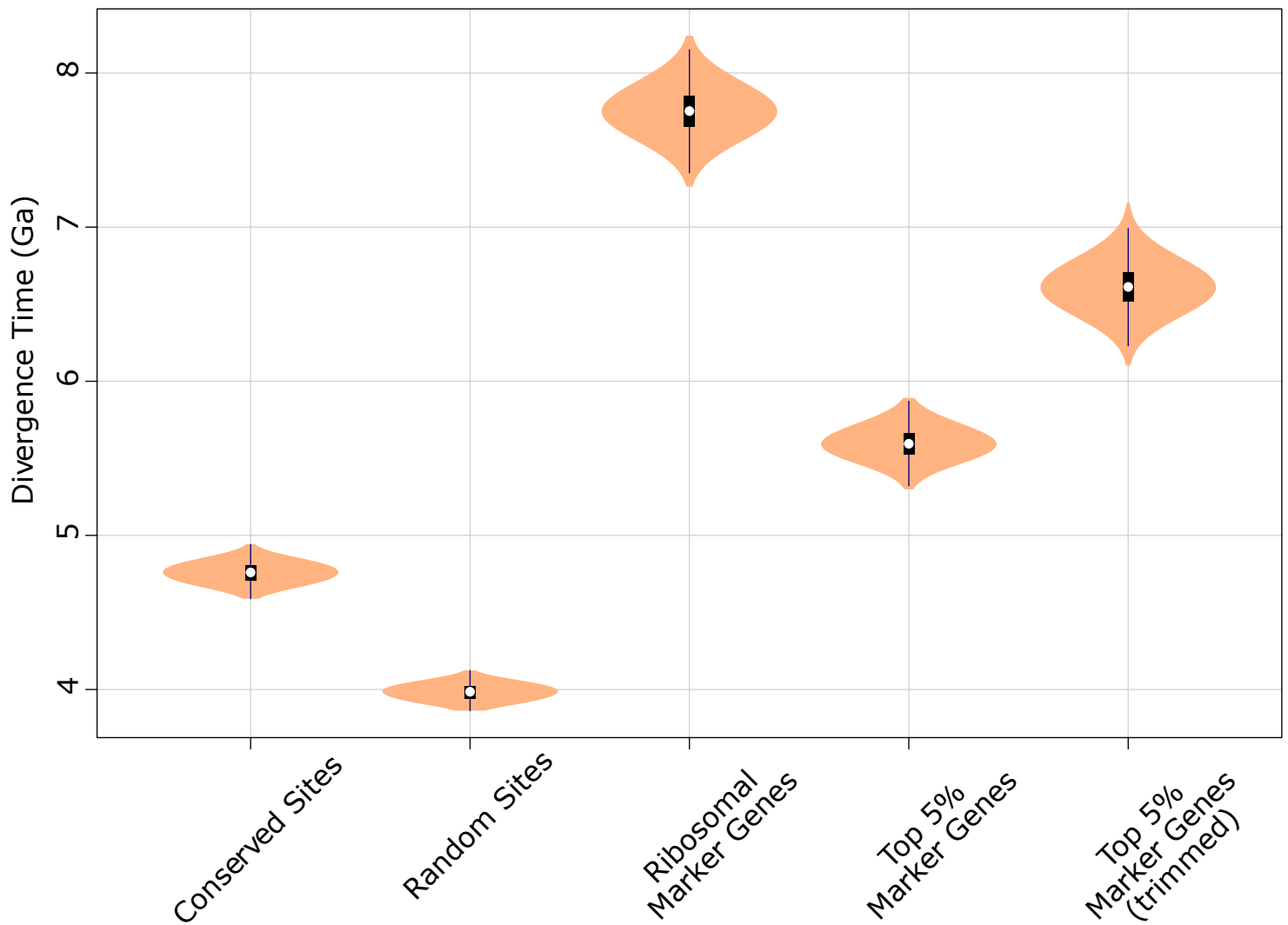


Figure 2. The inferred age of LUCA is not a reliable indicator of marker quality. Posterior node age estimates from Bayesian molecular clock analyses of 1) Conserved sites as estimated previously (Zhu et al., 2019); 2) Random sites (Zhu et al., 2019) 3) Ribosomal genes (Zhu et al., 2019) 4) The top 5% of marker gene families according to their ΔLL score (including only 1 ribosomal protein) and 5) The same top 5% of marker genes trimmed using BMGE (Criscuolo and Gribaldo, 2010) to remove highly variable sites. In each case, a strict molecular clock was applied, with the age of the Cyanobacteria-Melainabacteria split constrained between 2.5 and 2.6 Ga.

267 Phylogeny of Archaea and Bacteria using ancient vertically- 268 evolving genes

269

270 *Finding ancient vertically-evolving genes*

271

272 To estimate the AB branch length and the phylogeny of prokaryotes using a dataset that
273 resolves some of the issues identified above, we performed a meta-analysis of several
274 previous studies to identify a consensus set of vertically-evolving marker genes. We identified
275 unique markers from these analyses by reference to the COG ontology (Dombrowski et al.,
276 2020; Galperin et al., 2019), extracted homologous sequences from a representative sample
277 of 350 archaeal and 350 bacterial genomes, and performed iterative phylogenetics and
278 manual curation to obtain a set of 54 markers that recovered archaeal and bacterial monophyly
279 (see Methods). Subsequently, we ranked these 54 genes by the extent to which they
280 recovered established within-domain relationships using the split score, a criterion described
281 previously (Dombrowski et al., 2020) (see Methods) yielding a final set of 27 markers that were
282 used for inferring an updated universal species tree (see below). Marker genes that better
283 resolved relationships within each domain also supported a longer AB branch length (Figure
284 3).

285

286 *Distributions of AB branch lengths for ribosomal and non-ribosomal marker genes are* 287 *similar*

288

289 Traditional universal marker sets include many ribosomal proteins (Ciccarelli et al., 2006;
290 Fournier and Gogarten, 2010; Harris et al., 2003; Hug et al., 2016; Williams et al., 2020). If
291 ribosomal proteins experienced accelerated evolution during the divergence of Archaea and
292 Bacteria, this might lead to the inference of an artifactually long AB branch length (Petitjean et
293 al., 2014; Zhu et al., 2019). To investigate this, we plotted the inter-domain branch lengths for
294 the 38 and 16 ribosomal and non-ribosomal genes, respectively, comprising the 54 marker
295 genes set. We found no evidence that there was a longer AB branch associated with ribosomal
296 markers (Figure 4; mean AB branch length for ribosomal proteins 1.35, mean for non-
297 ribosomal 2.25). Prior to manual curation, non-ribosomal markers had a greater number of
298 HGTs and cases of mixed paralogy. In particular, for the original set of 95 markers, 62% of
299 the non-ribosomal markers and 21% of the ribosomal markers were not monophyletic,
300 respectively. These values were 69% and 29% for the 54 markers, and 50% and 33% for the
301 27 markers. These results imply that manual curation of marker genes is important for deep
302 phylogenetic analyses, particularly when using non-ribosomal markers.

303

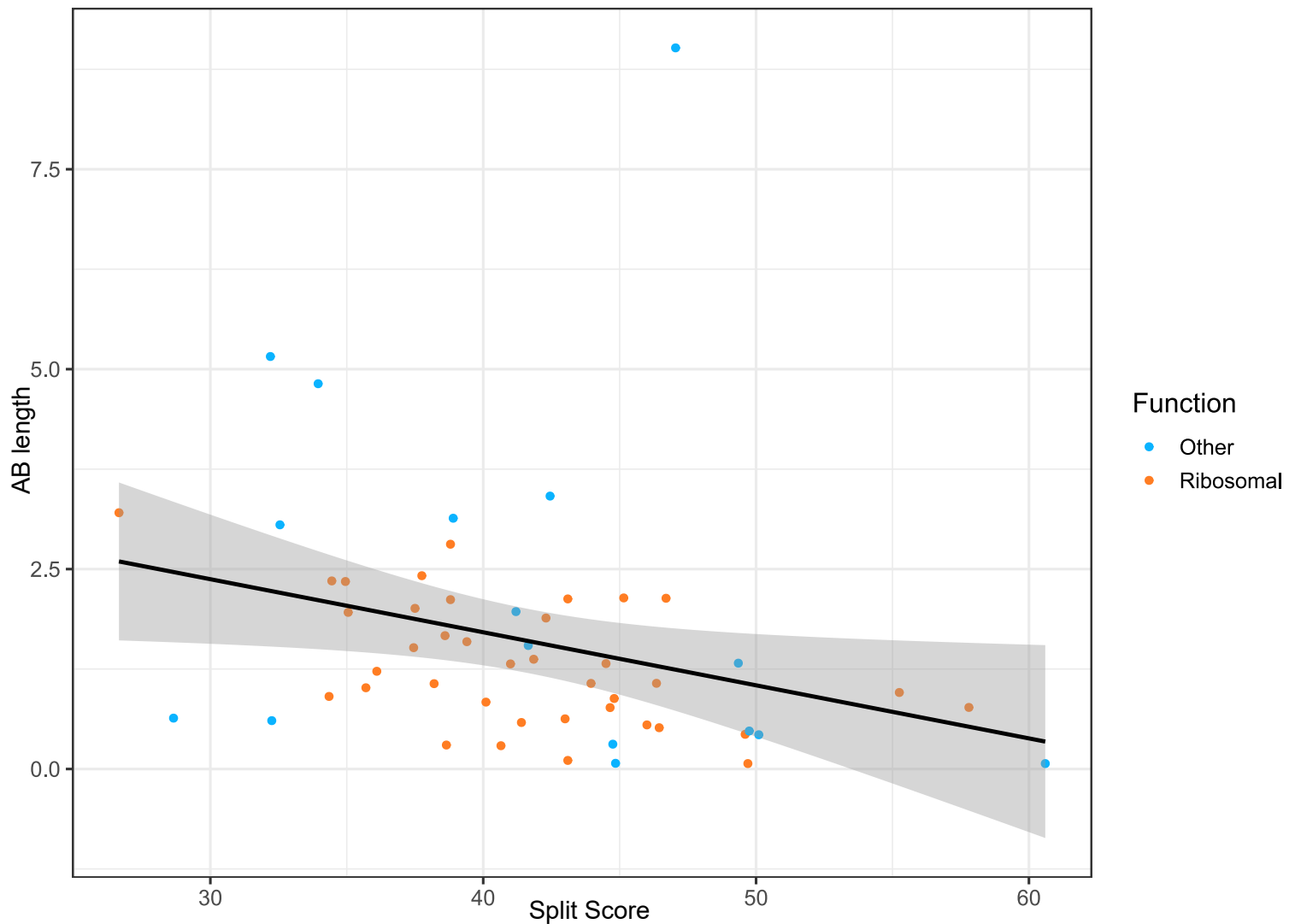


Figure 3. Better phylogenetic markers have longer AB branches. The plot shows the relationship between split score (a lower split score denotes better recovery of established within-domain relationships, see Methods) and AB branch length (in expected number of substitutions/site) for the 54 highest-ranked marker genes. Marker genes with higher split scores (that split monophyletic groups into multiple subclades) have shorter AB branch lengths ($P = 0.0311$, $r = 0.294$). Split scores of ribosomal and non-ribosomal markers were statistically indistinguishable ($P = 0.828$, Figure S3).

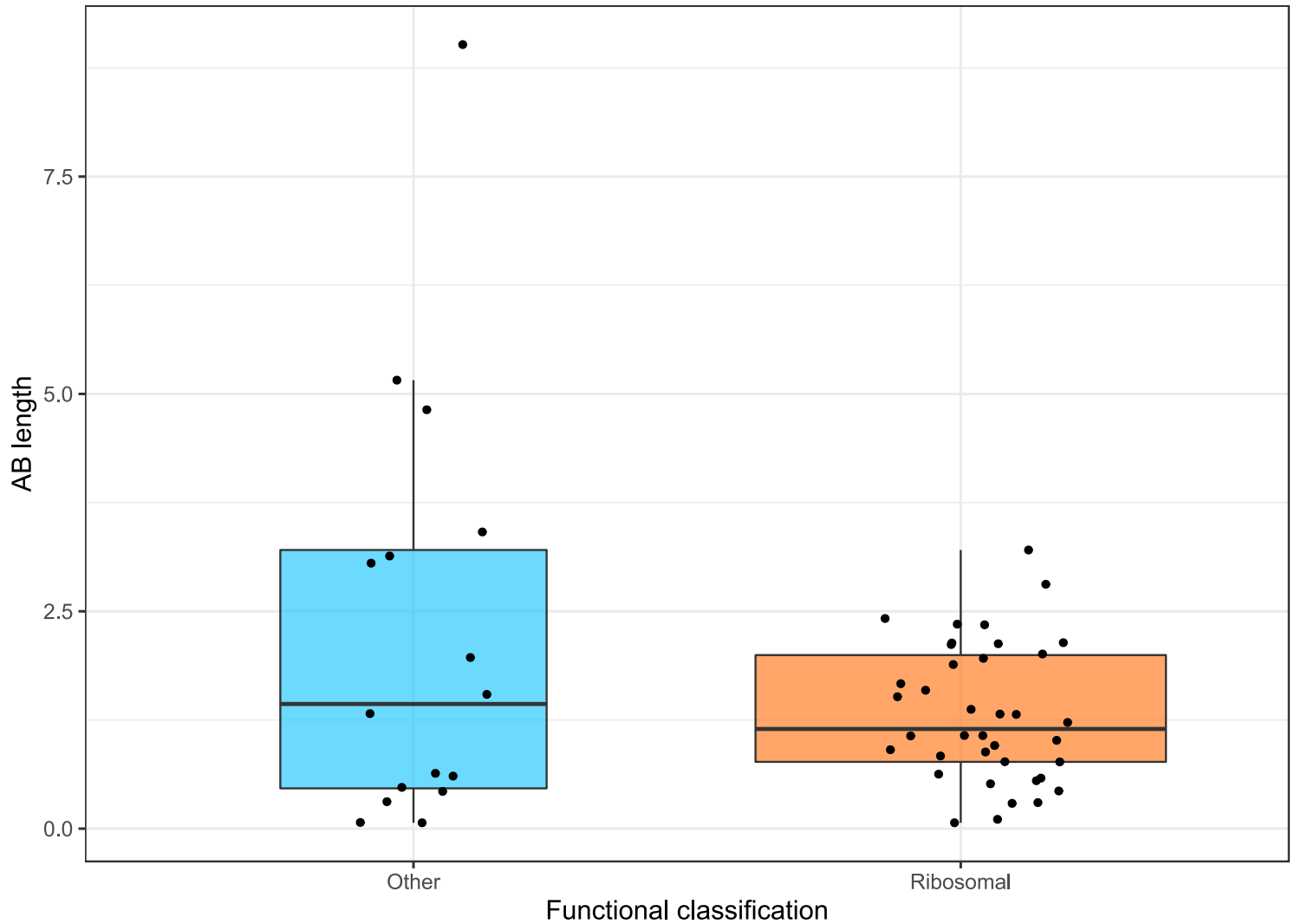


Figure 4. Among vertically-evolving marker genes, ribosomal genes do not have a longer AB branch length. The plot shows functional classification of markers against AB branch length using the 54 most vertically evolving markers. We did not see a significant ($P = 0.619$, Wilcoxon rank sum test) difference between AB branch lengths for ribosomal and non-ribosomal genes.

304 ***Substitutional saturation and poor model fit contribute to underestimation of AB branch***
305 ***length***

306

307 For the top 50% of marker genes as determined by split scores (27 genes), we performed an
308 additional round of single gene tree inference and manual review to identify and remove
309 remaining sequences which had evidence of HGT or represented distant paralogs. The
310 resulting single gene trees are provided in the Data Supplement
311 (10.6084/m9.figshare.13395470). To evaluate the relationship between site evolutionary rate
312 and AB branch length, we created two concatenations: fastest sites (comprising sites with
313 highest probability of being in the fastest Gamma rate category; 868 sites) and slowest sites
314 (sites with highest probability of being in the slowest Gamma rate category, 1604 sites) and
315 compared relative branch lengths inferred from the entire concatenate using IQ-TREE 2 to
316 infer site-specific rates (Figure 5). As expected, total tree length is shorter from the slow-
317 evolving sites, but the relative AB branch length is longer (1.2 substitutions/site, or ~2% of
318 total tree length, compared to 2.6 substitutions/site, or ~0.04% total tree length for the fastest-
319 evolving sites). This result suggests that, at fast-evolving sites, some changes along the AB
320 branch have been overwritten by later events in evolution --- that is, that substitutional
321 saturation leads to an underestimate of the AB branch length.

322

323 Another factor that has been shown to lead to underestimation of genetic distance on deep
324 branches is a failure to adequately model the site-specific features of sequence evolution
325 (Lartillot and Philippe, 2004; Schrempf et al., 2020; Wang et al., 2018; Williams et al., 2020).
326 Amino acid preferences vary across the sites of a sequence alignment, due to variation in the
327 underlying functional constraints (Lartillot and Philippe, 2004; Quang et al., 2008; Wang et al.,
328 2008). The consequence is that, at many alignment sites, only a subset of the twenty possible
329 amino acids are tolerated by selection. Standard substitution models, such as LG+G4+F, are
330 site-homogeneous, and approximate the composition of all sites using the average
331 composition across the entire alignment. Such models underestimate the rate of evolution at
332 highly constrained sites because they do not account for the high number of multiple
333 substitutions that occur at such sites. The effect is that site-homogeneous models
334 underestimate branch lengths when fit to site-heterogeneous data. Site-heterogeneous
335 models have been developed that account for site-specific amino acid preferences, and these
336 generally show improved fit to real protein sequence data (reviewed in (Williams et al., 2021)).
337 To evaluate the impact of substitution model fit for these data, we fit a range of models to the
338 full concatenation, assessing model fit using the Bayesian information criterion (BIC) in IQ-
339 TREE 2. The AB branch length inferred under the best-fit model, the site-heterogeneous
340 LG+C60+G4+F model, was 2.52 substitutions/site, ~1.7-fold greater than the branch length
341 inferred from the site-homogeneous LG+G4+F model (1.45 substitutions/site). Thus,
342 substitution model fit has a major effect on the estimated length of the AB branch, with better-
343 fitting models supporting a longer branch length (Table 1). The same trends are evident when
344 better-fitting site-heterogeneous models are used to analyse the dataset of Zhu et al.:
345 considering only the top 5% of genes by Δ LL score, the AB branch length is 1.2 under
346 LG+G4+F, but increases to 2.4 under the best-fitting LG+C60+G4+F model (Figure S2).

347

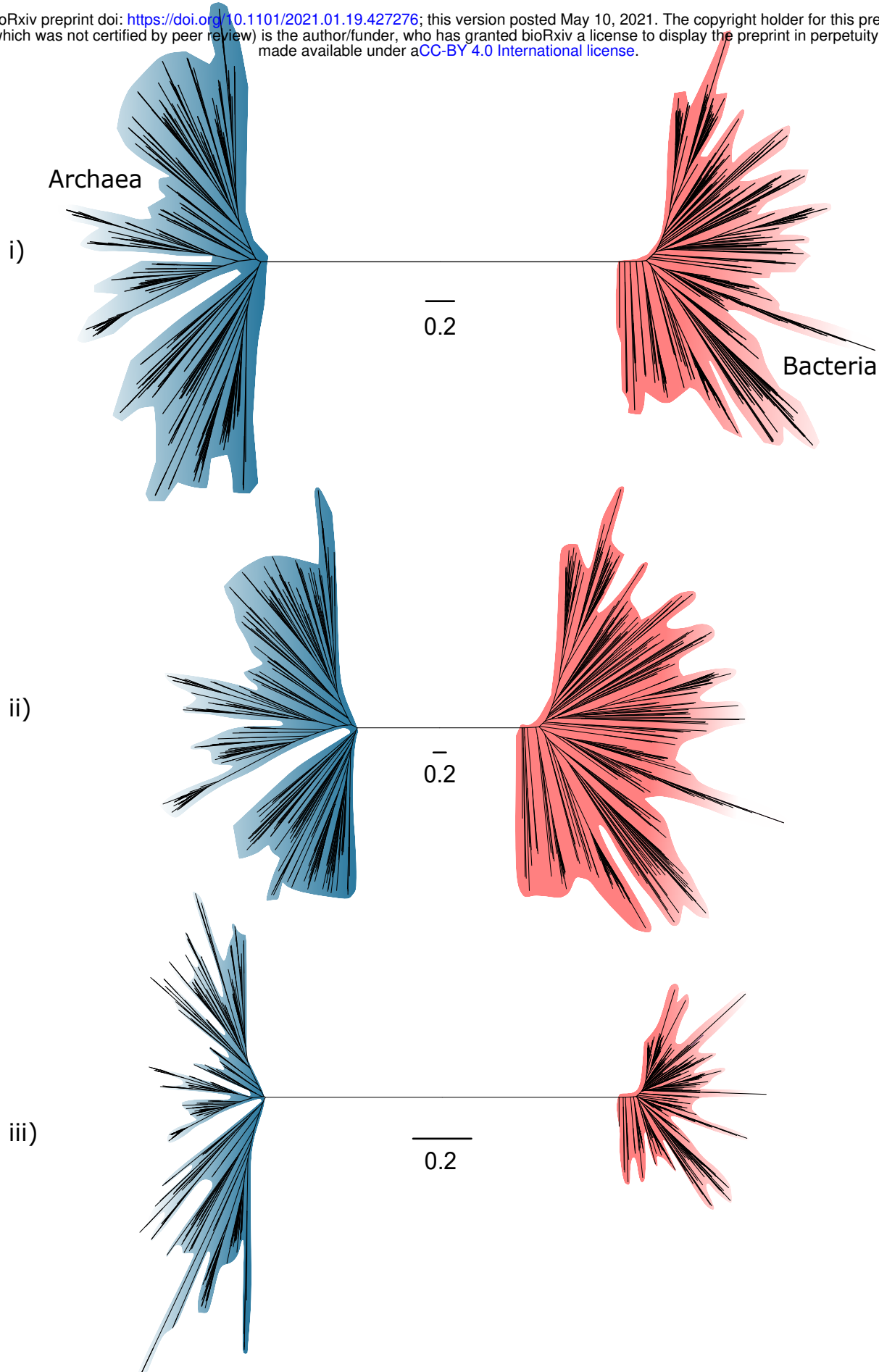


Figure 5. Slow- and fast-evolving sites support different shapes for the universal tree. (i) Tree of Archaea and Bacteria inferred from a concatenation of 27 core genes; (ii) Tree inferred from the fastest-evolving sites; (iii) Tree inferred from the slowest-evolving sites. To facilitate comparison of relative diversity, scale bars are provided separately for each panel. Slow-evolving sites support a relatively long inter-domain branch and less diversity within the domains (that is, shorter between-taxa branch lengths within domains). This suggests that substitution saturation (overwriting of earlier changes) may reduce the relative length of the AB branch at fast-evolving sites and genes.

348
349

Substitution model	BIC	AB branch length
LG+G4+F	5935950.053	1.449090256
LG+C20+G4+F	5783903.997	2.139350118
LG+C40+G4+F	5756823.360	2.469702112
LG+C60+G4+F	5746886.292	2.517828771

350 **Table 1. The inferred AB length from a concatenation of the top 27 markers using a**
351 **simple model versus models which account for site compositional heterogeneity.** Using
352 better fitting models, i.e models which allow for across-site compositional heterogeneity, a
353 longer AB branch is inferred.

354

355 ***A phylogeny of Archaea and Bacteria inferred from 27 vertically-evolving marker genes***

356

357 The topology of our phylogeny of the primary domains of life (Figure 6) is consistent with recent
358 single-domain trees inferred for Archaea and Bacteria independently (Coleman et al., 2021;
359 Dombrowski et al., 2020; Williams et al., 2017), although the deep relationships within Bacteria
360 are only poorly resolved, with the exception of the monophyly of Gracilicutes (Figure 6). A
361 recent analysis suggested that, among extant lineages, the metabolisms of Clostridia,
362 Deltaproteobacteria, Actinobacteria and some Aquificae might best preserve the metabolism
363 of the last bacterial common ancestor (Xavier et al., 2021). Assuming a universal root between
364 Archaea and Bacteria (Dagan et al., 2010; Gogarten et al., 1989; Iwabe et al., 1989), none of
365 these groups branch near the bacterial root in our analysis (Figure 6). This is consistent with
366 previous work (Castelle and Banfield, 2018; Hug et al., 2016; Parks et al., 2017; Raymann et
367 al., 2015) including the inference of an updated and rooted bacteria phylogeny (Coleman et
368 al., 2021). Notably, our analysis placed the Candidate Radiation (CPR) (Brown et al., 2015)
369 as a sister lineage to Chloroflexi (Chloroflexota) rather than as a deep-branching bacterial
370 superphylum. While this contrasts with initial trees suggesting that CPR may represent an
371 early diverging sister lineage of all other Bacteria (Brown et al., 2015; Castelle and Banfield,
372 2018; Hug et al., 2016), our finding is consistent with recent analyses that recovered CPR
373 within the Terrabacteria (Coleman et al., 2021; Taib et al., 2020). Together, these analyses
374 suggest that the deep-branching position of CPR in some trees was a result of long branch
375 attraction, a possibility that has been raised previously (Hug et al., 2016; Méheust et al., 2019).

376

377 The deep branches of the archaeal subtree are well-resolved in the ML tree and recover clades
378 of DPANN (albeit at 51% bootstrap support), Asgard (100% bootstrap support), and TACK
379 Archaea (75% bootstrap support), in agreement with a range of previous studies (Dombrowski
380 et al., 2020; Guy and Ettema, 2011; Raymann et al., 2015; Williams et al., 2017). We also find
381 support for the placement of Methanonatronarchaeia (Sorokin et al., 2017) distant to
382 Halobacteria within the Methanotecta, in agreement with recent analyses and suggesting their
383 initial placement with Halobacteria (Sorokin et al., 2017) may be an artifact of compositional
384 attraction (Aouad et al., 2019; Dombrowski et al., 2020; Martijn et al., 2020). Notably, the
385 Hadesarchaea (92% bootstrap support) and a clade comprising Theionarchaea,
386 Methanofastidiosa, and Thermococcales (92% bootstrap support) branch basal to the TACK

387 and Asgard Archaea, respectively, in our analysis, rather than with other Euryarchaeota.
388 These positions have been previously reported (Adam et al., 2017; Raymann et al., 2015;
389 Williams et al., 2017), though the extent of euryarchaeotal paraphyly and the lineages involved
390 has varied among analyses.

391

392 A broader observation from our analysis is that the phylogenetic diversity of the archaeal and
393 bacterial domains, measured as substitutions per site in this consensus set of vertically-
394 evolving marker genes, appears to be similar (Figure 5(i); the mean root to tip distance for
395 archaea: 2.38, for bacteria: 2.41, the range of root to tip distances for archaea: 1.79-3.01, for
396 bacteria: 1.70-3.17). Considering only the slowest-evolving category of sites, branch lengths
397 within Archaea are actually longer than within Bacteria (Figure 5(iii)). This result differs from
398 some published trees (Hug et al., 2016; Zhu et al., 2019) in which the phylogenetic diversity
399 of Bacteria has appeared to be significantly greater than that of Archaea. By contrast to those
400 earlier studies, we analysed a set of 350 genomes from each domain, an approach which may
401 tend to reduce the differences between them. While we had to significantly downsample the
402 sequenced diversity of Bacteria, our sampling nonetheless included representatives from all
403 known major lineages of both domains, and so might be expected to recover a difference in
404 diversity, if present. Our analyses and a number of previous studies (Hug et al., 2016; Parks
405 et al., 2018; Petitjean et al., 2014; Zhu et al., 2019) indicate that the choice of marker genes
406 has a profound impact on the apparent phylogenetic diversity of prokaryotic groups; for
407 instance, in the proportion of bacterial diversity composed of CPR (Hug et al., 2016; Parks et
408 al., 2017). Our results demonstrate that slow and fast-evolving sites from the same set of
409 marker genes support different tree shapes and branch lengths; it therefore seems possible
410 that between-dataset differences are due, at least in part, to evolutionary rate variation within
411 and between marker genes.

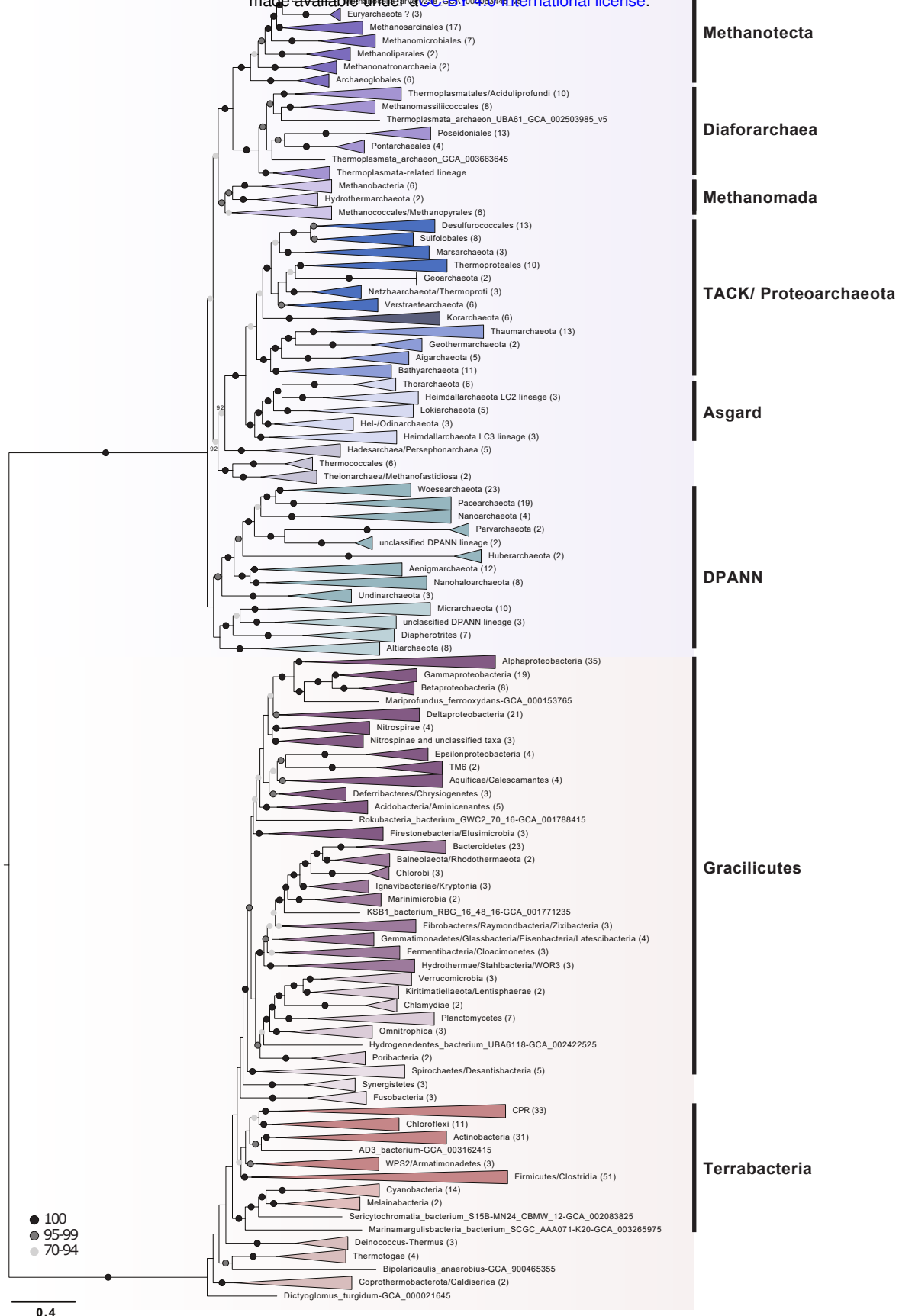


Figure 6: A phylogeny of Archaea and Bacteria inferred from a concatenation of 27 marker genes. Consistent with some recent studies (Dombrowski et al., 2020; Guy and Ettema, 2011; Raymann et al., 2015; Williams et al., 2017), we recovered the DPANN, TACK and Asgard Archaea as monophyletic groups. Although the deep branches within Bacteria are poorly resolved, we recovered a sister group relationship between CPR and Chloroflexota, consistent with a recent report (Coleman et al., 2021). The tree was inferred using the best-fitting LG+C60+G4+F model in IQ-TREE 2 (Minh et al., 2020). Branch lengths are proportional to the expected number of substitutions per site. Support values are ultrafast (UFBoot2) bootstraps (Hoang et al., 2018). Numbers in parenthesis refer to number of taxa within each collapsed clade. Please note that collapsed taxa in the Archaea and Bacteria roughly correspond to order- and phylum-level lineages, respectively.

412 Conclusion

413

414 Core gene phylogenies provide a window into the earliest period of archaeal and bacterial
415 evolution. Concatenation is useful for pooling signal across individual genes, but topology and
416 branch length estimates from concatenations only reflect the underlying tree of life if the
417 individual genes share the same evolutionary history. Our analysis of published datasets
418 (Coleman et al., 2021; Petitjean et al., 2014; Williams et al., 2020; Zhu et al., 2019) indicates
419 that incongruence among marker genes resulting from inter-domain gene transfer and hidden
420 paralogy can lead to an under-estimate of the inter-domain branch length. We performed a re-
421 analysis of marker genes from a range of published analyses, manually curated datasets to
422 identify and remove transferred genes, and estimated an updated phylogeny of Archaea and
423 Bacteria. Considering only this manually curated consensus marker gene dataset, we found
424 no evidence that ribosomal markers overestimate stem length; since they appear to be
425 transferred less frequently than other genes, our analysis affirms that ribosomal proteins are
426 useful markers for deep phylogeny. In general, better markers, regardless of functional
427 category, support a longer AB branch length. A phylogeny inferred from the 27 best-performing
428 markers was consistent with some recent work on early prokaryotic evolution, resolving the
429 major clades within Archaea and nesting the CPR within Terrabacteria. Our analyses suggest
430 that both the true Archaea-Bacteria branch length (Figure 7), and the phylogenetic diversity of
431 Archaea, may be underestimated by even the best current models, a finding that is consistent
432 with a root for the tree of life between the two prokaryotic domains.

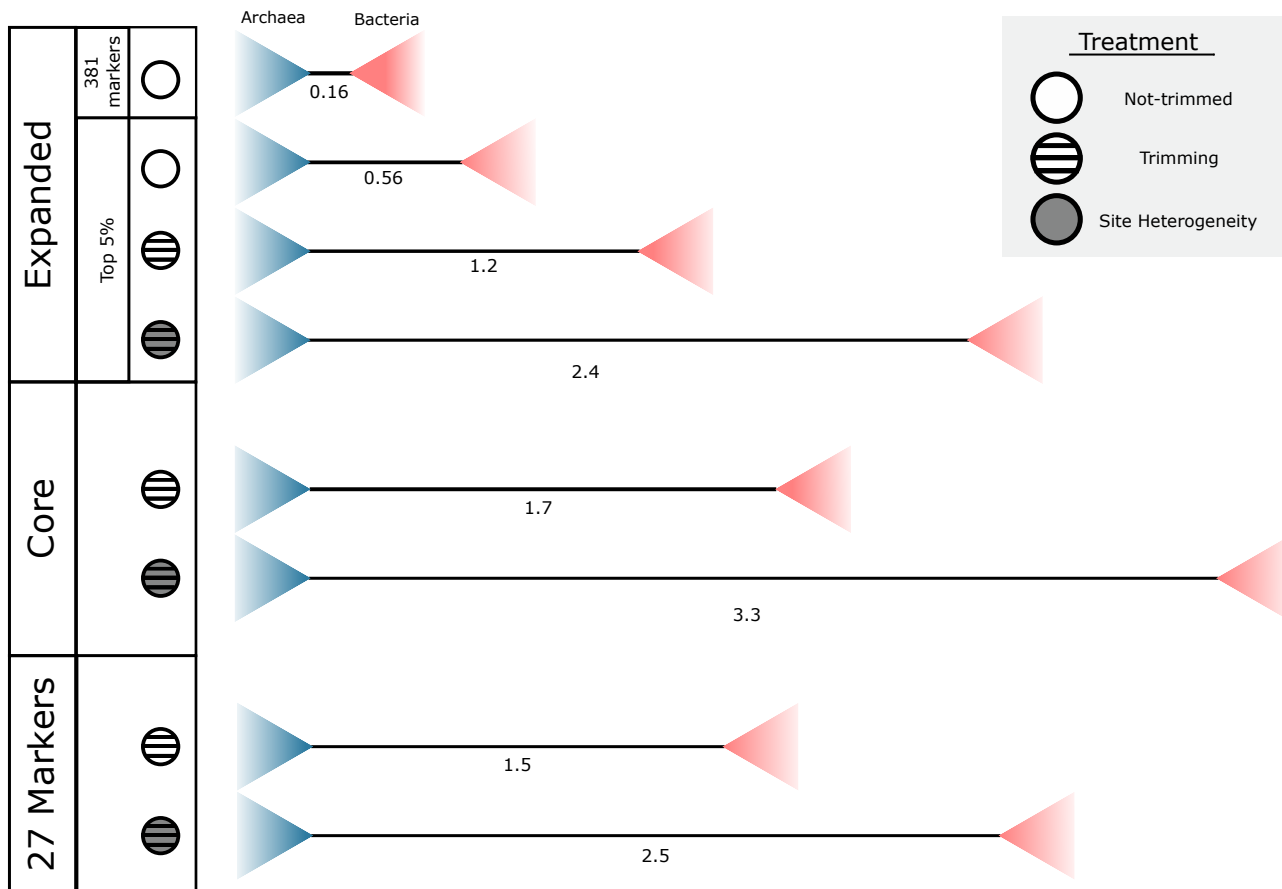


Figure 7. The impact of marker gene choice, phylogenetic congruence, alignment trimming, and substitution model fit on estimates of the Archaea-Bacteria branch length. Analysis using a site-homogeneous model (LG+G4+F) on the complete 381-gene expanded set results in an AB branch substantially shorter than previous estimates. Removing the genes most seriously affected by inter-domain gene transfer, trimming poorly-aligned sites using BMGE (Criscuolo and Gribaldo, 2010), and using the best-fitting site-heterogeneous model available (LG+C60+G4+F) substantially increase the estimated AB length, such that it is comparable with published estimates from the “core” set (Williams et al., 2020) and the consensus set of 27 markers identified in the present study. Branch lengths measured in expected number of substitutions/site.

433 Methods

434 **Data**

435 We downloaded the individual alignments from (Zhu et al., 2019)
436 (<https://github.com/biocore/wol/tree/master/data/>), along with the genome metadata and the
437 individual newick files. We checked each published tree for domain monophyly, and also
438 performed approximately unbiased (AU) (Shimodaira, 2002) tests to assess support for
439 domain monophyly on the underlying sequence alignments using IQ-TREE 2 (Minh et al.,
440 2020). The phylogenetic analyses were carried out using the ‘reduced’ subset of 1000 taxa
441 outlined by the authors (Zhu et al., 2019), for computational tractability. These markers were
442 also trimmed according to the protocol in the original paper (Zhu et al., 2019), i.e sites with
443 >90% gaps were removed, followed by removal of sequences with >66% gaps.

444 We also downloaded the Williams et al. (Williams et al., 2020) (“core”), Petitjean et al.
445 (Petitjean et al., 2014) (“non-ribosomal”) and Coleman et al. (Coleman et al., 2021)
446 (“bacterial”) datasets from their original publications.

447

448 **Annotations**

449 Proteins used for phylogenetic analyses by Zhu *et al.* (Zhu et al., 2019), were annotated to
450 investigate the selection of sequences comprising each of the marker gene families. To this
451 end, we downloaded the protein sequences provided by the authors from the following
452 repository: <https://github.com/biocore/wol/tree/master/data/alignments/genes>. To obtain
453 reliable annotations, we analysed all sequences per gene family using several published
454 databases, including the arCOGs (version from 2014) (Seemann, 2014), KOs from the KEGG
455 Automatic Annotation Server (KAAS; downloaded April 2019) (Aramaki et al., 2020), the Pfam
456 database (Release 31.0)(Bateman et al., 2004), the TIGRFAM database (Release 15.0) (Haft
457 et al., 2003), the Carbohydrate-Active enZymes (CAZy) database (downloaded from dbCAN2
458 in September 2019)(Cantarel et al., 2009), the MEROPs database (Release 12.0) (Rawlings
459 et al., 2016), (Saier et al., 2006), the hydrogenase database (HydDB; downloaded in
460 November 2018) (Søndergaard et al., 2016), the NCBI- non-redundant (nr) database
461 (downloaded in November 2018), and the NCBI COGs database (version from 2020).
462 Additionally, all proteins were scanned for protein domains using InterProScan (v5.29-68.0;
463 settings: --iprlookup --goterms) (Jones et al., 2014).

464

465 Individual database searches were conducted as follows: arCOGs were assigned using PSI-
466 BLAST v2.7.1+ (settings: -evalue 1e-4 -show_gis -outfmt 6 -max_target_seqs 1000 -dbsize
467 100000000 -comp_based_stats F -seg no) (Altschul et al., 1997). KOs (settings: -E 1e-5),
468 PFAMs (settings: -E 1e-10), TIGRFAMs (settings: -E 1e-20) and CAZymes (settings: -E 1e-
469 20) were identified in all archaeal genomes using hmmsearch v3.1b2(Finn et al., 2011). The
470 MEROPs and HydDB databases were searched using BLASTp v2.7.1 (settings: -outfmt 6, -
471 evalue 1e-20). Protein sequences were searched against the NCBI_nr database using
472 DIAMOND v0.9.22.123 (settings: -more-sensitive -e-value 1e-5 -seq 100 -no-self-hits -
473 taxonmap prot.accession2taxid.gz) (Buchfink et al., 2015). For all database searches the best
474 hit for each protein was selected based on the highest e-value and bitscore and all results are
475 summarized in the Data Supplement Table,
476 Annotation_Tables/0_Annotation_tables_full/All_Zhu_marker_annotations_16-12-

477 2020.tsv.zip. For InterProScan we report multiple hits corresponding to the individual domains
478 of a protein using a custom script (parse_IPRdomains_vs2_GO_2.py).

479
480 Assigned sequence annotations were summarized and all distinct KOs and Pfams were
481 collected and counted for each marker gene. KOs and Pfams with their corresponding
482 descriptions were mapped to the marker gene file downloaded from the repository:
483 <https://github.com/biocore/wol/blob/master/data/markers/metadata.xlsx> and used in
484 summarization of the 381 marker gene protein trees (Table S1).

485
486 For manual inspection of single marker gene trees, KO and Pfam annotations were mapped
487 to the tips of the published marker protein trees, downloaded from the repository:
488 <https://github.com/biocore/wol/tree/master/data/trees/genes>. Briefly, the Genome ID, Pfam,
489 Pfam description, KO, KO description, and NCBI Taxonomy string were collected from each
490 marker gene annotation table and were used to generate mapping files unique to each marker
491 gene phylogeny, which links the Genome ID to the annotation information
492 (GenomeID|Domain|Pfam|Pfam Description|KO|KO Description). An in-house perl script
493 `replace_tree_names.pl`
494 (https://github.com/ndombrowski/Phylogeny_tutorial/tree/main/Input_files/5_required_Scripts
495) was used to append the summarized protein annotations to the corresponding tips in each
496 marker gene tree. Annotated marker gene phylogenies were manually inspected using the
497 following criteria including: 1) retention of reciprocal domain monophyly (Archaea and
498 Bacteria) and 2) for the presence or absence of potential paralogous families. Paralogous
499 groups and misannotated families present in the gene trees were highlighted and violations of
500 search criteria were recorded in Table S1.

501 ***Phylogenetic analyses***

502 *COG assignment for the Core, Non-Ribosomal, and Bacterial marker genes*

503 First, all gene sequences in the three published marker sets (core, non-ribosomal, and
504 bacterial) were annotated using the NCBI COGs database (version from 2020). Sequences
505 were assigned a COG family using `hmmsearch v3.3.2` (Finn et al., 2011) (settings: `-E 1e-5`)
506 and the best hit for each protein sequence was selected based on the highest e-value and bit
507 score. To assign the appropriate COG family for each marker gene, we quantified the
508 percentage distribution of all unique COGs per gene, and selected the family representing the
509 majority of sequences in each marker gene.

510 Accounting for overlap, this resulted in 95 unique COG families from the original 119 total
511 marker genes across all three published datasets (Table S2). Orthologues corresponding to
512 these 95 COG families were identified in the 700 genomes (350 Archaea, 350 Bacteria, Table
513 S3) using `hmmsearch v3.3.2` (settings: `-E 1e-5`). The reported BinID and protein accession
514 were used to extract the sequences from the 700 genomes, which were used for subsequent
515 phylogenetic analyses.

516 *Marker gene inspection and analysis*

517 We aligned these 95 sequence sets using `MAFFT-linsi` (Katoh and Toh, 2008) and removed
518 poorly-aligned positions with `BMGE` (Criscuolo and Gribaldo, 2010). We inferred initial
519 maximum likelihood trees (LG+G4+F) for all 95 markers and mapped the KO and Pfam

520 domains and descriptions, inferred from annotation of the 700 genomes, to the corresponding
521 tips (see above). Manual inspection took into consideration monophyly of Archaea and
522 Bacteria and the presence of paralogs, and other signs of contamination (HGT, LBA).
523 Accordingly, single gene trees that failed to meet reciprocal domain monophyly were excluded,
524 and any instances of HGT, paralogous sequences, and LBA artefacts were manually removed
525 from the remaining trees resulting in 54 markers across the three published datasets that were
526 subject to subsequent phylogenetic analysis (LG+C20+G4+F) and further refinement (see
527 below).

528 529 *Ranking markers based on split score*

530 We applied an automated marker gene ranking procedure devised previously, the split score
531 (Dombrowski et al., 2020), to rank each of the 54 markers that satisfied reciprocal monophyly
532 based on the extent to which they recovered established phylum-, class- or, order-level
533 relationships within the archaeal and bacterial domains (Table S4).

534 The script quantifies the number of splits, or occurrences where a taxon fails to cluster within
535 its expected taxonomic lineage, across all gene phylogenies. Monophyly of archaeal and
536 bacterial lineages was assessed based on clades defined in Table S4. Briefly, we used
537 Cluster1 for Archaea in combination with Cluster0 (phylum) or Cluster3 (i.e. on class-level if
538 defined and otherwise on phylum-level; Table S4) for Bacteria. We then ranked the marker
539 genes using the following split-score criteria: the number of splits per taxon and the splits
540 normalized to the species count. The percentage of split phylogenetic groups was used to
541 determine the highest ranking (top 50%) markers.

542 *Concatenation*

543 Based on the split score ranking of the 54 marker genes (above), the top 50% (27 markers,
544 Table S4) marker genes were manually inspected using criteria as defined above, and
545 contaminating sequences were manually removed from the individual sequence files.
546 Following inspection, marker protein sequences were aligned using MAFFT-LINSI (Kato and
547 Standley, 2013) and trimmed using BMGE (Criscuolo and Gribaldo, 2010). We concatenated
548 the 27 markers into a supermatrix, which was used to infer a maximum-likelihood tree (Figure
549 6, under LG+C60+G4+F), evolutionary rates (see below), and rate-category supermatrices
550 as well as to perform model performance tests (see below).

551 *Constraint analysis*

552 We performed a maximum likelihood free topology search using IQ-TREE 2 (Minh et al., 2020)
553 under the LG+G4+F model, with 1000 bootstrap replicates on each of the markers from the
554 expanded, bacterial, core and non-ribosomal sets. We also performed a constrained analysis
555 with the same model, in order to find the maximum likelihood tree in which Archaea and
556 Bacteria were reciprocally monophyletic. We then compared both trees using the
557 approximately unbiased (AU) Shimodaira (2002) test in IQ-TREE 2 (Minh et al., 2020) with
558 10,000 RELL (Shimodaira, 2002) bootstrap replicates. To evaluate the relationship between
559 marker gene verticality and AB branch length, we calculated the difference in log-likelihood
560 between the constrained and unconstrained trees in order to rank the genes from the
561 expanded marker set, made concatenates comprised of the top 20-100 (intervals of 5) of these

562 marker genes, and inferred the tree length under LG+C10+G4+F with 1000 bootstrap
563 replicates.

564 *Site and gene evolutionary rates*

565 We inferred rates using the --rate option in IQ-TREE 2 (Minh et al., 2020) for both the 381
566 marker concatenation from Zhu (Zhu et al., 2019) and the top 5% of marker genes based on
567 the results of difference in log-likelihood between the constrained tree and free-tree search in
568 the constraint analysis (above). We also used this method to explore the differences in rates
569 for the 27 marker set. We built concatenates for sites in the slowest and fastest rate categories,
570 and inferred branch lengths from each of these concatenates using the tree inferred from the
571 corresponding dataset as a fixed topology.

572 *Substitution model fit*

573 Model fit tests were undertaken using the top 5% concatenate described above, with the
574 alignment being trimmed with BMGE 1.12 (Criscuolo and Gribaldo, 2010) with default settings
575 (BLOSUM62, entropy 0.5) for all of the analyses except the 'untrimmed' LG+G4+F run, other
576 models on the trimmed alignment were LG+G4+F, LG+R4+F and
577 LG+C10,20,30,40,50,60+G4+F, with 1000 bootstrap replicates. Model fitting was done using
578 ModelFinder (Kalyaanamoorthy et al., 2017) in IQ-TREE 2 (Minh et al., 2020). For the model
579 testing for the 27 concatenation, we performed a model finder analysis (-m MFP) including
580 additional complex models of evolution, (i.e.
581 LG+C60+G4+F, LG+C50+G4+F, LG+C40+G4+F, LG+C30+G4+F, LG+C20+G4+F, LG+C10+G
582 4+F, LG+G4+F, LG+R4+F) to the default, to find the best fitting model for the analysis. This
583 revealed that, according to AIC, BIC and cAIC, LG+C60+G4+F was the best fitting model. For
584 comparison, we also performed analyses using the following models:
585 LG+G4+F, LG+C20+G4+F, LG+C40+G4+F (Table 1).

586 *Molecular clock analyses*

587 Molecular clock analyses were devised to test the effect of genetic distance on the inferred
588 age of LUCA. Following the approach of Zhu et al (Zhu et al., 2019), we subsampled the
589 alignment to 100 species. Five alternative alignments were analysed, representing conserved
590 sites across the entire alignment, randomly selected sites across the entire alignment, only
591 ribosomal marker genes, the top 5% of marker genes according to Δ LL and the top 5% of
592 marker genes further trimmed under default settings in BMGE 1.12 (Criscuolo and Gribaldo,
593 2010). Divergence time analyses were performed in MCMCTree (Yang, 2007) under a strict
594 clock model. We used the normal approximation approach, with branch lengths estimated in
595 codeml under the LG+G4 model. In each case, a fixed tree topology was used alongside a
596 single calibration on the Cyanobacteria-Melainabacteria split. The calibration was modelled
597 as a uniform prior distribution between 2.5 and 2.6 Ga, with a 2.5% probability that either
598 bound could be exceeded. For each alignment, four independent MCMC chains were run for
599 2,000,000 generations to achieve convergence.

600 *Plotting*

601 Statistical analyses were performed using R 4.0.4 (R Core Team, 2021), and data were plotted
602 with ggplot2 (Wickham, 2016).

603 **Data and code availability**

604 All of the data, including sequence alignments, trees, annotation files, and scripts associated
605 with this manuscript have been deposited in the FigShare repository at DOI:
606 10.6084/m9.figshare.13395470.
607

608 **Acknowledgements**

609 ERRM was supported by a Royal Society Enhancement Award (RGF\EA\180199) to TAW.
610 CP was supported by NERC grant NE/P00251X/1 to TAW. This work was funded by the
611 Gordon and Betty Moore Foundation through grant GBMF9741 to TAW, AS and GJSz. TAW
612 was supported by a Royal Society University Research Fellowship (URF\R\201024). GJSz
613 received funding from the European Research Council under the European Union's Horizon
614 2020 research and innovation program under Grant Agreement 714774 and Grant GINOP-
615 2.3.2.-15-2016- 00057. AS is supported by the Swedish Research Council (VR starting grant
616 2016-03559), the NWO-I foundation of the Netherlands Organisation for Scientific Research
617 (WISE fellowship) and the European Research Council (ERC Starting grant 947317,
618 ASymbEL).

619

620 References

621

- 622 Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S. 2017. The growing tree of Archaea:
623 new perspectives on their diversity, evolution and ecology. *ISME J* **11**:2407–2425.
- 624 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.
625 Gapped BLAST and PSI-BLAST: a new generation of protein database search
626 programs. *Nucleic Acids Research* **25**:3389–3402.
- 627 Aouad M, Borrel G, Brochier-Armanet C, Gribaldo S. 2019. Evolutionary placement of
628 Methanonatronarchaeia. *Nature Microbiology* **4**:558–559.
- 629 Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020.
630 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score
631 threshold. *Bioinformatics* **36**:2251–2252.
- 632 Barboni M, Boehnke P, Keller B, Kohl IE, Schoene B, Young ED, McKeegan KD. 2017. Early
633 formation of the Moon 4.51 billion years ago. *Science Advances* **3**:e1602365.
- 634 Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M,
635 Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. 2004. The Pfam protein
636 families database. *Nucleic Acids Research* **32**:D138–41.
- 637 Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. 2018. Integrated
638 genomic and fossil evidence illuminates life's early evolution and eukaryote origin.
639 *Nature Ecology and Evolution* **2**:1556–1562.
- 640 Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC,
641 Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than
642 15% of domain Bacteria. *Nature* **523**:208–211.
- 643 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
644 *Nat Methods* **12**:59–60.
- 645 Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. The
646 Carbohydrate-Active EnZymes database (CAZy): an expert resource for
647 Glycogenomics. *Nucleic Acids Research* **37**:D233–8.
- 648 Castelle CJ, Banfield JF. 2018. Major New Microbial Groups Expand Diversity and Alter our
649 Understanding of the Tree of Life. *Cell* **172**:1181–1197.
- 650 Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic
651 reconstruction of a highly resolved tree of life. *Science* **311**:1283–1287.
- 652 Coleman GA, Davín AA, Mahendrarajah T, Spang A, Hugenholtz P, Szöllösi GJ, Williams
653 TA. 2021. A rooted phylogeny resolves early bacterial evolution. *Science* **372**.
654 doi:10.1126/science.abe5011
- 655 Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaeobacterial origin of
656 eukaryotes. *Proceedings of the National Academy of Sciences of the United States of*
657 *America* **105**:20356–20361.
- 658 Creevey CJ, Doerks T, Fitzpatrick DA, Raes J, Bork P. 2011. Universally distributed single-
659 copy genes indicate a constant rate of horizontal transfer. *PLoS One* **6**:e22099.
- 660 Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new
661 software for selection of phylogenetic informative regions from multiple sequence
662 alignments. *BMC Evolutionary Biology* **10**:210.
- 663 Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. 2017. Lokiarchaea are close relatives
664 of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS*
665 *Genetics* **13**:e1006810.
- 666 Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome networks root the tree of life
667 between prokaryotic domains. *Genome Biology and Evolution* **2**:379–392.
- 668 Dombrowski N, Williams TA, Sun J, Woodcroft BJ, Lee J-H, Minh BQ, Rinke C, Spang A.

- 669 2020. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on
670 archaeal evolution. *Nature Communications* **11**:1–15.
- 671 Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity
672 searching. *Nucleic Acids Research* **39**:W29–W37.
- 673 Foster PG. 2004. Modeling compositional heterogeneity. *Systematic Biology* **53**:485–495.
- 674 Fournier GP, Gogarten JP. 2010. Rooting the ribosomal tree of life. *Molecular Biology and*
675 *Evolution* **27**:1792–1801.
- 676 Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. 2019. Microbial genome
677 analysis: the COG approach. *Briefings in Bioinformatics* **20**:1063–1070.
- 678 Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ,
679 Date T, Oshima T, Konishi J, Denda K, Yoshida M. 1989. Evolution of the vacuolar H⁺-
680 ATPase: implications for the origin of eukaryotes. *Proceedings of the National Academy*
681 *of Sciences of the United States of America* **86**:6661–6665.
- 682 Gouy R, Baurain D, Philippe H. 2015. Rooting the tree of life: the phylogenetic jury is still out.
683 *Philosophical Transactions of the Royal Society B Biological Sciences* **370**:20140329.
- 684 Guy L, Ettema TJG. 2011. The archaeal “TACK” superphylum and the origin of eukaryotes.
685 *Trends in Microbiology* **19**:580–587.
- 686 Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. *Nucleic*
687 *Acids Research* **31**:371–373.
- 688 Hanan BB, Tilton GR. 1987. 60025: relict of primitive lunar crust? *Earth and Planetary*
689 *Science Letters* **84**:15–21.
- 690 Harris JK, Kelley ST, Spiegelman GB, Pace NR. 2003. The genetic core of the universal
691 ancestor. *Genome Research* **13**:407–412.
- 692 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the
693 Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**:518–522.
- 694 Horita J, Berndt ME. 1999. Abiogenic methane formation and isotopic fractionation under
695 hydrothermal conditions. *Science* **285**:1055–1057.
- 696 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
697 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,
698 Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nature*
699 *Microbiology* **1**:16048 doi:10.1038/nmicrobiol.2016.48
- 700 Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of
701 archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of
702 duplicated genes. *Proceedings of the National Academy of Sciences of the United*
703 *States of America* **86**:9355–9359.
- 704 Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of
705 incongruence? *Trends in Genetics* **22**:225–231.
- 706 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell
707 A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y,
708 Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification.
709 *Bioinformatics* **30**:1236–1240.
- 710 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermin LS. 2017. ModelFinder:
711 fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587–589.
- 712 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
713 improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–
714 780.
- 715 Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment
716 program. *Briefings in Bioinformatics*. **9**(4):286-298 doi:10.1093/bib/bbn013
- 717 Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in
718 the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*
719 **7**:S4.
- 720 Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the
721 amino-acid replacement process. *Molecular Biology and Evolution* **21**:1095–1109.
- 722 Lepland A, Arrhenius G, Cornell D. 2002. Apatite in early Archean Isua supracrustal rocks,
723 southern West Greenland: its origin, association with graphite and potential as a

- 724 biomarker. *Precambrian Research* **118**:221–241.
- 725 Martijn J, Schön ME, Lind AE, Vosseberg J, Williams TA, Spang A, Ettema TJG. 2020.
- 726 Hikarchaea demonstrate an intermediate stage in the methanogen-to-halophile
- 727 transition. *Nature Communications* **11**:5490.
- 728 Méheust R, Burstein D, Castelle CJ, Banfield JF. 2019. The distinction of CPR bacteria from
- 729 other bacteria based on protein family content. *Nature Communications* **10**:4173.
- 730 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
- 731 R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
- 732 the Genomic Era. *Molecular Biology and Evolution* **37**:1530–1534.
- 733 Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL:
- 734 genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**:i541–8.
- 735 Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Göker M,
- 736 Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, Yoshikuni Y, Visel A,
- 737 Whitman WB, Garrity GM, Eisen JA, Hugenholtz P, Pati A, Ivanova NN, Woyke T, Klenk
- 738 H-P, Kyrpides NC. 2017. 1,003 reference genomes of bacterial and archaeal isolates
- 739 expand coverage of the tree of life. *Nature Biotechnology* **35**:676–683.
- 740 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz
- 741 P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially
- 742 revises the tree of life. *Nature Biotechnology* **36**: 996-1004. doi:10.1038/nbt.4229
- 743 Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P,
- 744 Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes
- 745 substantially expands the tree of life. *Nature Microbiology* **2**:1533–1542.
- 746 Petitjean C, Deschamps P, López-García P, Moreira D. 2014. Rooting the domain archaea
- 747 by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota.
- 748 *Genome Biology and Evolution* **7**:191–204.
- 749 Pühler G, Leffers H, Gropp F, Palm P, Klenk HP, Lottspeich F, Garrett RA, Zillig W. 1989.
- 750 Archaeobacterial DNA-dependent RNA polymerases testify to the evolution of the
- 751 eukaryotic nuclear genome. *Proceedings of the National Academy of Sciences of the*
- 752 *United States of America* **86**:4569–4573.
- 753 Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic
- 754 reconstruction. *Bioinformatics* **24**:2317–2323.
- 755 Ramulu HG, Groussin M, Talla E, Planel R, Daubin V, Brochier-Armanet C. 2014. Ribosomal
- 756 proteins: toward a next generation standard for prokaryotic systematics? *Molecular*
- 757 *Phylogenetics and Evolution* **75**:103–117.
- 758 Rawlings ND, Barrett AJ, Finn R. 2016. Twenty years of the MEROPS database of
- 759 proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research* **44**:D343–
- 760 50.
- 761 Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a
- 762 new root for the Archaea. *Proceedings of the National Academy of Sciences* **112**:6670–
- 763 6675.
- 764 R Core Team. 2021. R: A language and environment for statistical computing. R Foundation
- 765 for Statistical Computing, Vienna, Austria.
- 766 Saier MH Jr, Tran CV, Barabote RD. 2006. TCDB: the Transporter Classification Database
- 767 for membrane transport protein analyses and information. *Nucleic Acids Research*
- 768 **34**:D181–6.
- 769 Schrempf D, Lartillot N, Szöllösi G. 2020. Scalable empirical mixture models that account for
- 770 across-site compositional heterogeneity. *Mol Biol Evol*. doi:10.1093/molbev/msaa145
- 771 Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068–
- 772 2069.
- 773 Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for
- 774 improved phylogenetic and taxonomic placement of microbes. *Nature Communications*
- 775 **4**:2304.
- 776 Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection.
- 777 *Systematic Biology* **51**:492–508.
- 778 Søndergaard D, Pedersen CNS, Greening C. 2016. HydDB: A web tool for hydrogenase

- 779 classification and analysis. *Scientific Reports* **6**:34212.
- 780 Sorokin DY, Makarova KS, Abbas B, Ferrer M, Golyshin PN, Galinski EA, Ciordia S, Mena
781 MC, Merkel AY, Wolf YI, van Loosdrecht MCM, Koonin EV. 2017. Discovery of
782 extremely halophilic, methyl-reducing euryarchaea provides insights into the
783 evolutionary origin of methanogenesis. *Nature Microbiology* **2**:17081.
- 784 Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R,
785 Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between
786 prokaryotes and eukaryotes. *Nature* **521**:173–179.
- 787 Sugitani K, Mimura K, Takeuchi M, Lepot K, Ito S. 2015. Early evolution of large micro-
788 organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-
789 walled microfossils. *Geobiology* **13**:507–521.
- 790 Taib, N, Megrian D, Witwinowski J, Adam P, Poppleton D, Borrel G, Beloin C, Gribaldo S.
791 2020. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm
792 transition. *Nature ecology & evolution*, **4**(12):1661-1672.
- 793 Tourasse NJ, Gouy M. 1999. Accounting for Evolutionary Rate Variation among Sequence
794 Sites Consistently Changes Universal Phylogenies Deduced from rRNA and Protein-
795 Coding Genes. *Molecular Phylogenetics and Evolution* **13**:159–168.
- 796 Valas RE, Bourne PE. 2011. The origin of a derived superkingdom: how a gram-positive
797 bacterium crossed the desert to become an archaeon. *Biology Direct* **6**:16.
- 798 van Zuilen MA, Lepland A, Arrhenius G. 2002. Reassessing the evidence for the earliest
799 traces of life. *Nature* **418**:627–630.
- 800 Wang H-C, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for
801 site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC*
802 *Evolutionary Biology* **8**:331.
- 803 Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling Site Heterogeneity with Posterior
804 Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.
805 *Systematic Biology* **67**:216–235.
- 806 Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.
- 807 Williams T a., Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic
808 signal places eukaryotes within the Archaea. *Philosophical Transactions of the Royal*
809 *Society B Biological Sciences* **279**:4870–4879.
- 810 Williams TA, Cox CJ, Foster PG, Szöllösi GJ, Embley TM. 2020. Phylogenomics provides
811 robust support for a two-domains tree of life. *Nature Ecology and Evolution* **4**:138–147.
- 812 Williams TA, Schrepf D, Szöllösi GJ, Cox CJ, Foster PG, Embley TM. 2021. Inferring the
813 deep past from molecular data. *Genome Biology and Evolution*.
814 doi:10.1093/gbe/evab067
- 815 Williams TA, Szöllösi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley
816 TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of
817 life. *Proceedings of the National Academy of Sciences of the United States of America*
818 **114**:E4602–E4611.
- 819 Xavier JC, Gerhards RE, Wimmer JLE, Brueckner J, Tria FDK, Martin WF. 2021. The
820 metabolic network of the last bacterial common ancestor. *Communications Biology* **4**:
821 413. doi:10.1038/s42003-021-01918-4
- 822 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and*
823 *Evolution* **24**:1586–1591.
- 824 Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA,
825 Kopylova E, McDonald D, Kosciölek T, Yin JB, Huang S, Salam N, Jiao J-Y, Wu Z, Xu
826 ZZ, Cantrell K, Yang Y, Sayyari E, Rabiee M, Morton JT, Podell S, Knights D, Li W-J,
827 Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. 2019. Phylogenomics of
828 10,575 genomes reveals evolutionary proximity between domains Bacteria and
829 Archaea. *Nature Communications* **10**. doi:10.1038/s41467-019-13443-4