Universal markers support a long inter-domain branch between Archaea and Bacteria

Edmund R. R. Moody[1], Tara A. Mahendrarajah[2], Nina Dombrowski[2], James W. Clark[1], Celine Petitjean[1], Pierre Offre[2], Gergely J. Szollosi[3,4,5], Anja Spang[2,6], Tom A. Williams[1*]

1. School of Biological Sciences, University of Bristol, Bristol BS8 1TQ, UK.
2. NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry; AB Den Burg, The Netherlands
3. Dept. of Biological Physics, Eötvös Loránd University, 1117 Budapest, Hungary
4. MTA-ELTE "Lendület" Evolutionary Genomics Research Group, 1117 Budapest, Hungary;
5. Evolutionary Systems Research Group, Centre for Ecological Research, Hungarian Academy of Sciences, 8237 Tihany, Hungary
6. Department of Cell- and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123, Uppsala, Sweden

*To whom correspondence should be addressed: tom.a.williams@bristol.ac.uk

Abstract

The tree of life is generally estimated from a core set of 16-56 genes coding for proteins predominantly involved in translation and other conserved informational and cellular processes. These markers represent only a fraction of the genes that were likely present in the last universal common ancestor (LUCA), but are useful for deep phylogenetic reconstructions because their mode of inheritance appears to be mainly vertical, which satisfies the assumptions of gene concatenation and supertree methods. Previous phylogenetic analyses of these genes recovered a long branch between Archaea and Bacteria. By contrast, a recent study made use of a greatly expanded set of 381 marker genes and recovered a much shorter branch length between Archaea and Bacteria, comparable to some divergences within the domains. These analyses suggest that the apparent deep split between Archaea and Bacteria may be the result of accelerated evolution of ribosomal genes. Here we re-evaluate the evolutionary history of the expanded marker gene set and show that substitutional saturation, inter-domain gene transfer, hidden paralogy, and poor model fit contribute to the inference of an artificially shortened inter-domain branch. Our results do not exclude a moderately faster rate of ribosomal gene evolution during the divergence of Archaea and Bacteria, but indicate that vertically-evolving marker genes across all functional categories support a major genetic divergence between the two primary domains of life.

Introduction

Much remains unknown about the earliest period of cellular evolution and the deepest divergences in the tree of life. Phylogenies encompassing both Archaea and Bacteria have been inferred from a "universal core" set of 16-56 genes encoding proteins involved in translation and other aspects of the genetic information processing machinery[1–10]. These genes are thought to evolve vertically and so more closely track an underlying tree of life compared to other genes[3,4,8,11]. In these analyses, the branch separating Archaea from Bacteria (hereafter, the AB branch) is often the longest internal branch in the tree[2,6,12–15]. In molecular phylogenetics, branch lengths are usually measured in expected numbers of substitutions per site, with a long branch corresponding to a greater degree of genetic change. Long branches can therefore result from high evolutionary rates, long periods of time, or a combination of the two. Molecular clock models can, in principle, disentangle the contributions of these effects, but only very few fossil calibrations[16] are currently available that are old enough to calibrate early divergences[17–20], and as a result the ages and evolutionary rates of the deepest branches of the tree remain highly uncertain.

Recently, Zhu et al.[21] inferred a phylogeny from 381 genes distributed across Archaea and Bacteria using the supertree method ASTRAL[22]. In addition to a large increase in the number of genes compared to other universal marker sets, the functional profile of these markers comprises not only proteins involved in information processing but also proteins affiliated with most other functional COG categories, including metabolic processes (Table S1). Subsequently, the genetic distance (branch length) between the domains[21] was estimated from a concatenation of the same marker genes, resulting in a much shorter AB branch length than observed with the core universal markers[2,6]. These analyses motivated the hypothesis[21] that the apparent deep divergence of Archaea and Bacteria might be the result of an accelerated evolutionary rate of genes encoding ribosomal proteins along the AB branch as compared to other genes. Interestingly, the same observation was made previously using a smaller set of 38 non-ribosomal marker proteins[5], although the difference in AB branch length between ribosomal and non-ribosomal markers in that analysis was reported to be substantially lower (roughly two-fold, compared to roughly ten-fold for the 381 protein set of Zhu et al.)[5,21].

A higher evolutionary rate of ribosomal genes might result from the accumulation of compensatory substitutions at the interaction surfaces among the protein subunits of the ribosome[5,23], or as a compensatory response to the addition or removal of ribosomal subunits early in evolution[5]. Alternatively, differences in the inferred AB branch length might result from varying rates or patterns of evolution between the traditional core genes[2,24] and the expanded set[21]. Substitutional saturation (multiple substitutions at the same site[25]) and across-site compositional heterogeneity can both impact the inference of tree topologies and branch lengths[26–30]. These difficulties are particularly significant for ancient divergences[31]. Failure to model site-specific amino acid preferences has previously been shown to lead to under-estimation of the AB branch length due to a failure to detect convergent changes [2,32], although the published analysis of the 381 marker set did not find evidence of a substantial impact of these features on the tree as a whole[21]. Those analyses also identified phylogenetic incongruence among the 381 markers, but did not determine the underlying cause[21].

This recent work[21] raises two important issues regarding the inference of the universal tree: first, that estimates of the genetic distance between Archaea and Bacteria from classic "core genes" may be unrepresentative of ancient genomes as a whole, and second, that there may be many more suitable genes to investigate early evolutionary history than generally recognized, providing an opportunity to improve the precision and accuracy of deep phylogenies. Here, we address these points by examining the evolutionary history of the 381 marker set (hereafter, the expanded marker gene set), and by evaluating the impact of orthology assignment, horizontal gene transfer (HGT), substitutional saturation and substitution model fit on inferences of the genetic divergence between Archaea and Bacteria based on gene concatenations.

Results and Discussion

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

The 381-gene set was originally derived from a larger set of 400 genes used to estimate the phylogenetic placement of new bacterial lineages as part of the PhyloPhlAn method[33]. Perhaps reflecting the focus on bacterial phylogeny in the original application, the phylogenetic distribution of the 381 marker genes in the expanded set varies substantially (Table S1), with many being poorly represented in Archaea. Indeed 25% of the published gene trees (https://biocore.github.io/wol/[21]) contain less than 0.5% archaeal homologues, with 21 (5%) and 69 (18%) of these trees containing no or less than 10 archaeal homologues, respectively. For the remaining 75% of the gene trees, archaeal homologs comprise 0.5%-13.4% of the dataset. While there are many more sequenced bacteria than archaea, 63% of the gene trees possessed genes from less than half of the 669 archaeal genomes included in the analysis , whereas only 22% of the gene trees possessed fewer than half of the total number of 9906 sampled bacterial genomes. These distributions suggest that many of these genes are not broadly present in both domains, and that some might be specific to Bacteria.

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

In the focal analysis of the 381 gene set, the tree topology was inferred using the supertree method ASTRAL[22], with branch lengths inferred on this fixed tree from a marker gene concatenation[21]. The topology inferred from this expanded marker set[21] is similar to published trees[6,34] and recovers Archaea and Bacteria as reciprocally monophyletic domains, albeit with a shorter AB branch than in earlier analyses. However, the individual gene trees[21] disagree regarding domain monophyly: Archaea and Bacteria are recovered as reciprocally monophyletic groups in only 24 of the 381 published[21] (Table S1) maximum likelihood (ML) gene trees for the expanded marker set.

Since single gene trees often fail to strongly resolve ancient relationships, we used approximately-unbiased (AU) tests[35] to evaluate whether the failure to recover domain monophyly in the published ML trees is statistically supported. For computational tractability,

we performed these analyses on a 1000-species subsample of the full dataset that was compiled in the original study[21]. For 79 of the 381 genes, we could not perform the test because the gene was not found on any of the 74 archaeal genomes present in the 1000-species subsample. For the remaining 302 genes, domain monophyly was rejected ($p < 0.05$) for 232 out of 302 (76.8%) genes. As a comparison, we performed the same test on several smaller marker sets used previously to infer a tree of life[2,5,36]; none of the markers in those sets rejected reciprocal domain monophyly (AU > 0.05 for all genes, Figure 1(a)). In what follows, we refer to four published marker gene sets: the Expanded set (381 genes[21]), the Core set (49 genes[2], encoding ribosomal proteins and other conserved information-processing functions), the Non-ribosomal set (38 genes, broadly distributed and explicitly selected to avoid genes encoding ribosomal proteins[5]), and the Bacterial set (29 genes used in a recent analysis of bacterial phylogeny[36]).

To investigate why 232 of the marker genes rejected the reciprocal monophyly of Archaea and Bacteria, we returned to the full 10,575-species dataset and annotated each sequence in each marker gene family by assigning proteins to KOs, Pfams, and Interpro domains, among others (Table S1, see Methods for details). We labelled the tips on each published marker gene phylogeny using the corresponding KO and PFAM annotations and descriptions (See Methods) and manually inspected the tree topologies for reciprocal domain monophyly of Archaea and Bacteria and the presence of paralogues (Table S1). This revealed that the major cause of domain polyphyly observed in gene trees was inter-domain gene transfer (in 357 out of 381 gene trees (93.7%)) and mixing of sequences from distinct paralogous families (in 246 out of 381 gene trees (64.6%)). For instance, marker genes encoding ABC-type transporters (p0131, p0151, p0159, p0174, p0181, p0287, p0306, po0364), tRNA synthetases (i.e. p0000, p0011, p0020, p0091, p0094, p0202), aminotransferases and dehydratases ((i.e. p0073/4-aminobutyrate aminotransferase; p0093/3-isopropylmalate dehydratase) often comprised a mixture of paralogues. For example, the phylogenetic tree comprising spermidine/putrescine import ATP-binding protein PotA (p0131) contains several paralogues families, such as different sugar or ion ATP-binding proteins. Further, dTDP-glucose 4,6-dehydratase (p0134) is representative of a gene tree that includes several paralogous families with different substrate specificities that include dehydrorhamnose and glucuronate.

Together, these analyses indicate that the evolutionary histories of the individual markers of the expanded set differ from each other and from the species tree. Zhu et al. acknowledged[21] the varying levels of congruence between the marker phylogenies and the species tree, but did not investigate the underlying causes. Our analyses establish the basis for these disagreements in terms of gene transfers and the mixing of orthologues and paralogues within and between domains. In principle, concatenation is based on the assumption that all of the genes in the supermatrix evolve on the same underlying tree; genes with different gene tree topologies should not be concatenated because the topological differences among sites are not modelled, and so the impact on inferred branch lengths is difficult to predict. In practice, it is often difficult to be certain that all of the markers in a concatenate share the same gene tree topology, and the analysis proceeds on the hypothesis that a small proportion of discordant genes are not expected to seriously impact the inferred tree. However, the concatenated tree inferred from the expanded marker set differs from previous trees in that the genetic distance between Bacteria and Archaea is greatly reduced, such that the AB branch length appears comparable to the distance between bacterial phyla[21]. An accurate estimate of the AB branch length is important because it has a major bearing on unanswered questions regarding the

root of the universal tree[31] and the deepest divisions among cellular life. We therefore evaluated the impact of the conflicting gene histories within the expanded marker set on inferred AB branch length.

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

To investigate the impact of gene transfers and mixed paralogy on the AB branch length inferred by gene concatenations[21], we compared branch lengths estimated from markers that rejected (AU < 0.05) or did not reject (AU > 0.05) the reciprocal monophyly of Bacteria and Archaea in the 381 marker set (Figure 1(a)). To estimate AB branch lengths for genes in which the domains were not monophyletic in the ML tree, we first performed a constrained ML search to find the best gene tree that was consistent with domain monophyly for each family under the LG+G4+F model in IQ-TREE 2[37]. While it may seem strained to estimate the length of a branch that does not appear in the ML tree, we reasoned that this approach would provide insight into the contribution of these genes to the AB branch length in the concatenation, in which they conflict with the overall topology. AB branch lengths were significantly (P = 2.159×10^{-12}, Wilcoxon rank sum test) shorter for markers that rejected domain monophyly (Figure 1(a); <0.05: mean AB branch length in expected substitutions/site 0.0130, >0.05: mean AB branch length 0.559). This result suggests that inter-domain gene transfers reduce the AB branch length when included in a concatenation. This behaviour might result from marker gene transfers reducing the number of fixed differences between the domains, so that the AB branch length in a tree in which Archaea and Bacteria are constrained to be reciprocally monophyletic will tend towards 0 as the number of transfers increases. Consistent with this hypothesis, we observed that ΔLL, the difference in log likelihood between the ML gene tree and the constrained ML tree consistent with domain monophyly (a simple proxy for marker gene verticality) correlates negatively with AB branch length (Figure 1(b)). Furthermore, AB branch length decreased as increasing numbers of low-verticality markers were added to the concatenate (Figure 1(c)). Taken together, these results indicate that the inclusion of genes that do not support the reciprocal monophyly of Archaea and Bacteria in the universal concatenate reduces the estimated AB branch length by homogenizing the genetic diversity of the two domains.

The inferred Archaea-Bacteria branch is artifactually shortened by unaccounted-for substitutional saturation

The longer AB branch length observed in concatenations of traditional core marker sets[2,6,24] compared to the expanded marker set[21] has been interpreted as evidence for ribosomal gene-specific accelerated evolution between domains[5,21], because ribosomal proteins and proteins that physically interact with the ribosome make up a large proportion (47.1%) of the core set[2]. An alternative hypothesis is that faster rates of evolution in the genes of the expanded set following the divergence of Archaea and Bacteria has resulted in substitutional saturation, overwriting some of the early changes and shortening the AB branch. To distinguish between these hypotheses, we compared AB branch lengths estimated from fast- and slow-evolving sites in the expanded concatenate, for the 1000 taxa subset[21]. We estimated site-specific rates of evolution using IQ-TREE 2[37], and constructed two new concatenates comprising the 25% fastest- and 25% slowest-evolving sites. As expected, the total tree length inferred from the slowest sites was shorter (105 substitutions/site compared to 1216 substitutions/site in the

25% fastest site concatenation). However, the relative length of the AB branch (that is, AB branch length as a proportion of total tree length) was 8.8-fold higher in the concatenation inferred from the 25% slowest sites compared to the fastest sites (Figure 2(a)). Considering only the top 5% of the expanded marker set in terms of **Δ**LL score, the relative difference increases to 11.1-fold. This analysis suggests that substitutional saturation has overwritten many of the changes that originally occurred during the divergence of Archaea and Bacteria at fast-evolving sites in the expanded gene set. Total tree lengths estimated for the expanded marker genes did not significantly differ (P = 0.5716) between genes that did or did not reject domain monophyly.

An earlier study[5] also recovered a longer AB branch from ribosomal compared to non-ribosomal genes, and none of the markers used in that study rejected domain monophyly (Figure 1, "Non-ribosomal"). While the non-ribosomal markers from that study[5] have similar total tree lengths to those in the core gene set (Figure 3(a)), they have ~2.4-fold shorter AB branches, Figure 3(b)). These results are consistent with the hypothesis that ribosomal proteins may have longer AB branches than other marker genes.

Failure to model across-site amino acid preferences reduces estimates of the AB branch length

Amino acid preferences vary across the sites of a sequence alignment, due to variation in the underlying functional constraints[26,29,30]. The consequence is that, at many alignment sites, only a subset of the twenty possible amino acids are tolerated by selection. Standard substitution models, such as LG+G+F, are site-homogeneous, and approximate the composition of all sites using the average composition across the entire alignment. Such models underestimate the rate of evolution at highly constrained sites because they cannot account for the high number of multiple substitutions that occur at such sites. The effect is that site-homogeneous models underestimate branch lengths when fit to site-heterogeneous data. The AB branch has previously been shown to be particularly susceptible to this effect[2]. Zhu et al.[21] investigated the impact of site heterogeneity on tree topology and branch lengths using the posterior mean site frequency (PMSF) model[38], and found that inferences under site-homogeneous and site-heterogeneous models were similar over the dataset as a whole[21]. To evaluate the effect of modelling site-heterogeneity on the AB branch length in particular, we inferred phylogenies from the concatenate constructed from the top 5% of genes scored by ΔLL under a series of models that account for site-heterogeneity using increasingly rich mixtures of site-specific composition profiles (C10-C60)[30]. Models that account for site heterogeneity support a branch that is 1.6-2.1-fold longer than using a site-homogeneous model; the best-fitting model among those we tested was LG+C60+G+F according to the BIC score, and this model inferred an AB branch length of 2.4; that is, twice as long as that inferred under the site-homogeneous model. Therefore, and consistent with previous work[2,32], substitution model fit has a significant effect on the inferred genetic proximity of Archaea and Bacteria (Figure 4). We also find that trimming poorly-aligned sites greatly increases the inferred length of the AB branch, likely because trimmed sites tend to be fast-evolving.

The age of the last universal common ancestor (LUCA) inferred from the expanded gene set is driven by variation in the AB branch length

Zhu et al.[21] argued that the expanded marker set is useful for deep phylogeny because estimates of the age of LUCA obtained by fitting molecular clocks to their dataset are in agreement with the geological record: a root (LUCA) age of 3.6-4.2 Gyr was inferred from the entire 381-gene dataset, consistent with the earliest fossil evidence for life, whereas estimates from ribosomal markers alone supported a root age of 7 Gya. This age might be considered unrealistic because it is much older than the age of the Earth (with the moon-forming impact occurring ~4.51Gya[42,43]). In the original analyses, the age of LUCA was estimated using a maximum likelihood approach, as well as a Bayesian molecular clock with a strict clock (assuming a constant evolutionary rate) or a relaxed clock with a single calibration. A strict clock model does not permit changes in evolutionary rate through time or across branches, and so a longer AB branch will lead to an older inferred LUCA age. Likewise, a relaxed clock model with a single calibration may fail to differentiate molecular distances and geological time. Given that the short AB branch in the expanded gene set results, in part, from phylogenetic incongruence among markers, we evaluated the age of LUCA inferred from the subset of the expanded gene set least affected by these issues. To do so, we analysed the top 5% of gene families according to their **Δ**LL score (including only 1 ribosomal protein) under the same clock model parameters as the original dataset. This analysis resulted in a significantly more ancient age estimate for LUCA (5.5-6.5 Ga), suggesting that, under these conditions, the inferred age for LUCA is driven by variation in the AB branch length, and is not in itself a reliable indicator of marker quality. In principle, more reliable estimates of LUCA's age might be obtained by using more calibrations. However, unambiguous calibrations remain elusive, particularly for the root and other deep branches of the tree. Despite advances in molecular clock methodology, such calibrations represent the only way to reliably capture the relationship between genetic distance and divergence time.

Conclusion

The apparent genetic proximity of Archaea and Bacteria inferred from a concatenation of the expanded gene set of Zhu et al. [21] results from substitutional saturation and inter-domain HGT. Saturation obscures evidence of substitutions along the AB branch, while undetected HGT acts to artifactually homogenize the genetic diversity of the domains, and leads to a reduction in the inferred AB branch when branch lengths are estimated on a fixed topology. Treatments of the data that improve estimation of the AB branch length - the use of better-fitting substitution models, or use of the markers least affected by inter-domain HGT - results in the inference of a ~15-fold longer AB branch that is similar to estimates from previously published datasets (Figure 6). These results emphasise the importance of the fit of the substitution model to the data, and the need to evaluate whether the data meet the assumptions of the methods used - in particular, the assumption that all genes evolve on the same underlying tree topology, which underlies analyses of gene concatenations. The violation of that assumption in branch length estimates from concatenated alignments of the expanded marker genes resulted in the recovery of an artifactually short AB branch.

A distinct question is whether the long AB branch inferred from congruent marker gene sets might be affected by accelerated ribosomal gene evolution. Comparison of AB branch length for sets of congruent "core" and non-ribosomal marker sets (Figure 3) are consistent with a higher relative rate of evolution along the AB branch for ribosomal genes, although the effect is small by comparison with the impact of phylogenetic congruence, site trimming and substitution model fit (Figure 6). In sum, all of our analyses support the conclusion that the split separating Archaea from Bacteria is by far the longest internal branch on the tree of the primary domains of life.

In the future, it may be useful to further evaluate the suitability of non-ribosomal markers for concatenated gene phylogenies. Furthermore, methods that explicitly model HGTs[44–46], while also calculating branch lengths in a manner that accounts for differences among the underlying gene trees, and that incorporate alternative means of dating information, such as gene transfers[17,47–49], may help to better constrain the divergence times of the deepest cellular lineages.

Methods

Data

We downloaded the individual alignments from [21] (https://github.com/biocore/wol/tree/master/data/), along with the genome metadata and the individual newick files. We checked each published tree for domain monophyly, and also performed approximately unbiased (AU)[35] tests to assess support for domain monophyly on the underlying sequence alignments using IQ-TREE 2[37]. The phylogenetic analyses were carried out using the 'reduced' subset of 1000 taxa outlined by the authors[21], for computational tractability. These markers were also trimmed according to the protocol in the original paper[21], i.e sites with >90% gaps were removed, followed by removal of sequences with >66% gaps.

We also downloaded the Williams et al.[2] ("core"), Petitjean et al.[5] ("non-ribosomal") and Coleman et al.[36] ("bacterial") datasets from their original publications.

Annotations

Proteins used for phylogenetic analyses by Zhu *et al*.[21], were annotated to investigate the selection of sequences comprising each of the marker gene families. To this end, we downloaded the protein sequences provided by the authors from the following repository: https://github.com/biocore/wol/tree/master/data/alignments/genes. To obtain reliable annotations, we analysed all sequences per gene family using several published databases, including the arCOGs (version from 2014)[50], KOs from the KEGG Automatic Annotation Server (KAAS; downloaded April 2019)[51], the Pfam database (Release 31.0)[52], the TIGRFAM database (Release 15.0)[53], the Carbohydrate-Active enZymes (CAZy) database (downloaded from dbCAN2 in September 2019)[54], the MEROPs database (Release 12.0)[55,56], the hydrogenase database (HydDB; downloaded in November 2018)[57], the NCBI- non-redundant (nr) database (downloaded in November 2018), and the NCBI COGs database (version from 2020). Additionally, all proteins were scanned for protein domains using InterProScan (v5.29-68.0; settings: --iprlookup --goterms)[58].

Individual database searches were conducted as follows: arCOGs were assigned using PSI-BLAST v2.7.1+ (settings: -evalue 1e-4 -show_gis -outfmt 6 -max_target_seqs 1000 -dbsize 100000000 -comp_based_stats F -seg no)[59]. KOs (settings: -E 1e-5), PFAMs (settings: -E 1e-10), TIGRFAMs (settings: -E 1e-20) and CAZymes (settings: -E 1e-20) were identified in all archaeal genomes using hmmsearch v3.1b2[60]. The MEROPsand HydDB databases were searched using BLASTp v2.7.1 (settings: -outfmt 6, -evalue 1e-20). Protein sequences were searched against the NCBI_nr database using DIAMOND v0.9.22.123 (settings: –more-sensitive –e-value 1e-5 –seq 100 –no-self-hits –taxonmap prot.accession2taxid.gz)[61]. For all database searches the best hit for each protein was selected based on the highest e-value and bitscore and all results are summarized in the Data Supplement, Annotation_Tables/0_Annotation_tables_full/All_Zhu_marker_annotations_16-12-2020.tsv.zip. For InterProScan we report multiple hits corresponding to the individual domains of a protein using a custom script (parse_IPRdomains_vs2_GO_2.py).

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

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

Phylogenetic analyses

Constraint analysis

We performed a maximum likelihood free topology search using IQ-TREE 2[37] under the LG+G+F model, with 10,000 bootstrap replicates. We also performed a constrained analysis with the same model, in order to find the maximum likelihood tree in which Archaea and Bacteria were reciprocally monophyletic. We then compared both trees using the approximately unbiased (AU)[35] test in IQ-TREE 2[37] with 10,000 RELL[35] bootstrap replicates. To evaluate the relationship between marker gene verticality and AB branch length, we calculated the difference in log-likelihood between the constrained and unconstrained trees in order to rank the genes from the expanded marker set, made concatenates comprised of the

top 20-100 (intervals of 5) of these marker genes, and inferred the tree length under LG+C10+G+F with 1000 bootstrap replicates.

Site and gene evolutionary rates

We inferred rates using the --rate option in IQ-TREE 2[37] for both the 381 marker concatenation from Zhu[21] and the top 5% of marker genes based on the results of difference in log-likelihood between the constrained tree and free-tree search in the constraint analysis (above). We built concatenates for the 25% slowest and 25% fastest sites, and inferred branch lengths from each of these concatenates using the tree inferred from the complete dataset as a fixed topology.

Individual markers

We inferred rates and trees (1000 bootstrap replicates) for each individual marker from Petitjean *et al.*[5], Williams *et al.*[2], Zhu *et al.*[21] under the LG+G+F and LG+C20+G+F models using IQ-TREE 2[37]. Mean rates per sequence were then calculated using a python script (see Data Supplement).

Model complexity

Model complexity tests were undertaken using the top 5% concatenate described above, with the alignment being trimmed with BMGE 1.12[41] with default settings (BLOSUM62, entropy 0.5) for all of the analyses except the 'untrimmed' LG+G+F run, other models on the trimmed alignment were LG+G+F, LG+R+F and LG+C10,20,30,40,50,60+G+F, with 1000 bootstrap replicates. Model fitting was done using ModelFinder[62] in IQ-TREE 2[37].

Molecular clock analyses

Molecular clock analyses were devised to test the effect of genetic distance on the inferred age of LUCA. Following the approach of Zhu et al[21], we subsampled the alignment to 100 species. Five alternative alignments were analysed, representing conserved sites across the entire alignment, randomly selected sites across the entire alignment, only ribosomal marker genes, the top 5% of marker genes according to ΔLL and the top 5% of marker genes further trimmed under default settings in BMGE 1.12[41]. Divergence time analyses were performed in MCMCTree[63] under a strict clock model. We used the normal approximation approach, with branch lengths estimated in codeml under the LG+G4 model. In each case, a fixed tree topology was used alongside a single calibration on the Cyanobacteria-Melainabacteria split. The calibration was modelled as a uniform prior distribution between 2.5 and 2.6 Ga, with a 2.5% probability that either bound could be exceeded. For each alignment, four independent MCMC chains were run for 2,000,000 generations to achieve convergence.

Plotting

Statistical analyses were performed using R 3.6.3[64], and data were plotted with ggplot2[65].

Figure legends

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 (AU P < 0.05) 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 (AU p > 0.05 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 a LOESS regression.

Figure 2: Vertically-evolving genes and slow-evolving sites support a longer relative AB branch length. We estimated site-specific evolutionary rates for all marker genes in the expanded dataset (A-B), as well as for the 20 genes with the smallest ΔLL (top 5%) in that dataset (C-D). Concatenations based on the 25% slowest sites (A,C) and on the top 5% vertical genes (C,D) support a longer AB branch. This suggests that the inference of a short AB branch is impacted by both substitutional saturation and unmodelled inter-domain transfer of marker genes. Phylogenies were inferred under the LG+G4+F model in IQ-TREE 2[37]. Branch lengths are the expected number of substitutions per site, as indicated by the scale bars. Alignment lengths in amino acids: **A**: 36797, **B**: 67274, **C**: 2736, **D**: 3884.

Figure 3. Evidence for a faster rate of inter-domain ribosomal protein evolution. (A) Tree lengths (total number of substitutions/site for each gene family) for the core and non-ribosomal gene sets under the LG+G+F and LG+C20+G+F models. Modelling site heterogeneity increases tree lengths (the number of inferred substitutions) due to improved modelling of the site-specific features of the evolutionary process. Tree lengths are not significantly different between the core and non-ribosomal marker sets P = 0.4821/0.1651 (for LG+G+F and LG+C60+G+F respectively) (B) Non-ribosomal genes have moderately shorter AB branch lengths than core genes, consistent with a moderately faster rate of ribosomal gene evolution on the inter-domain branch. The difference is significant under both LG+G+F ($P = 8.78 * 10^{-9}$), and the better-fitting LG+C20+G+F model ($P = 2.237 * 10^{-7}$).

Figure 4. The effect of modelling site heterogeneity on AB branch length. Increasing the number of protein mixture profiles, as well as trimming is associated with a change in AB branch length on the expanded marker set[21]. All analyses used LG exchangeabilities, four rate categories (Gamma-distributed or freely estimated), and included a general composition vector containing the empirical amino acid frequencies (+F). Modelling of site heterogeneity with the C10-C60 models increases the inferred AB branch length ~2-fold. Trimming poorly-aligned sites slightly increases the AB branch estimation whereas relaxing the gamma rate categories slightly decreases estimation of AB branch length. LG (LG substitution matrix), G (four gamma rate categories), F (empirical site frequencies estimated from the data), C10-60 (number of protein mixture profiles used[30]) R (four free rate categories which relax the assumption of a gamma distribution for rates[39,40], BMGE (trimming using Block Mapping and Gathering with Entropy[41]). The trendline is estimated using a LOESS regression.

Figure 5. Divergence time estimation of the Archaea-Bacteria split. Violins represent posterior age estimates from Bayesian molecular clock analyses of 1) Conserved sites as estimated previously[21]; 2) Random sites[21] 3) Ribosomal genes[21] 4) The top 5% of marker gene families according to their **Δ**LL score (including only 1 ribosomal protein) and 5) The same top 5% of marker genes trimmed using BMGE[41] 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.

Figure 6. 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+G+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[41], and using the best-fitting site-heterogeneous model available (LG+C60+G+F) substantially increase the estimated AB length, such that it is comparable with published estimates from non-ribosomal and core gene sets. Branch lengths measured in expected number of substitutions/site.

Data and code availability

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

Acknowledgements

ERRM was supported by a Royal Society Enhancement Award (RGF\EA\180199) to TAW. CP was supported by NERC grant NE/P00251X/1 to TAW. TAW was supported by a Royal Society University Research Fellowship (URF\R\201024). GJSz received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program under Grant Agreement 714774 and Grant GINOP-2.3.2.-15-2016-00057. AS was supported by the Swedish Research Council (VR starting grant 2016-03559

524 to AS) and the NWO-I foundation of the Netherlands Organisation for Scientific Research
525 (WISE fellowship to AS).
526
527
528
529

References

1. Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).
2. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).
3. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
4. Ramulu, H. G. *et al.* Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.* **75**, 103–117 (2014).
5. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. *Genome Biol. Evol.* **7**, 191–204 (2014).
6. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* vol. 1 (2016).
7. Fournier, G. P. & Gogarten, J. P. Rooting the ribosomal tree of life. *Mol. Biol. Evol.* **27**, 1792–1801 (2010).
8. Theobald, D. L. A formal test of the theory of universal common ancestry. *Nature* **465**, 219–222 (2010).
9. Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).
10. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences* **112**, 6670–6675 (2015).
11. Creevey, C. J., Doerks, T., Fitzpatrick, D. A., Raes, J. & Bork, P. Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One* **6**, e22099 (2011).
12. Gogarten, J. P. *et al.* Evolution of the vacuolar H+-ATPase: implications for the origin of

557 eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 6661–6665 (1989).
558 13. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. Evolutionary relationship of
559 archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of
560 duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 9355–9359 (1989).
561 14. Pühler, G. *et al.* Archaebacterial DNA-dependent RNA polymerases testify to the
562 evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 4569–
563 4573 (1989).
564 15. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaebacterial
565 origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).
566 16. Sugitani, K., Mimura, K., Takeuchi, M., Lepot, K. & Ito, S. Early evolution of large micro-
567 organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-
568 walled microfossils. *geobiology* **13**, 507–521 (2015).
569 17. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early
570 evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
571 18. van Zuilen, M. A., Lepland, A. & Arrhenius, G. Reassessing the evidence for the earliest
572 traces of life. *Nature* **418**, 627–630 (2002).
573 19. Horita, J. & Berndt, M. E. Abiogenic methane formation and isotopic fractionation under
574 hydrothermal conditions. *Science* **285**, 1055–1057 (1999).
575 20. Lepland, A., Arrhenius, G. & Cornell, D. Apatite in early Archean Isua supracrustal
576 rocks, southern West Greenland: its origin, association with graphite and potential as a
577 biomarker. *Precambrian Res.* **118**, 221–241 (2002).
578 21. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity
579 between domains Bacteria and Archaea. *Nature Communications* **10,** 5477 (2019).
580 22. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation.
581 *Bioinformatics* **30**, i541–8 (2014).
582 23. Valas, R. E. & Bourne, P. E. The origin of a derived superkingdom: how a gram-positive
583 bacterium crossed the desert to become an archaeon. *Biol. Direct* **6**, 16 (2011).
584 24. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and

585 eukaryotes. *Nature* **521**, 173–179 (2015).

586 25. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of
587 incongruence? *Trends Genet.* **22**, 225–231 (2006).

588 26. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in
589 the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

590 27. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

591 28. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts
592 in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**, S4
593 (2007).

594 29. Wang, H.-C., Li, K., Susko, E. & Roger, A. J. A class frequency mixture model that
595 adjusts for site-specific amino acid frequencies and improves inference of protein
596 phylogeny. *BMC Evolutionary Biology* **8**, 331 (2008).

597 30. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for
598 phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).

599 31. Gouy, R., Baurain, D. & Philippe, H. Rooting the tree of life: the phylogenetic jury is still
600 out. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140329 (2015).

601 32. Tourasse, N. J. & Gouy, M. Accounting for Evolutionary Rate Variation among
602 Sequence Sites Consistently Changes Universal Phylogenies Deduced from rRNA and
603 Protein-Coding Genes. *Molecular Phylogenetics and Evolution* **13**, 159–168 (1999).

604 33. Segata, N., Börnigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method
605 for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**,
606 2304 (2013).

607 34. Castelle, C. J. & Banfield, J. F. Major New Microbial Groups Expand Diversity and Alter
608 our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

609 35. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst.*
610 *Biol.* **51**, 492–508 (2002).

611 36. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Cold Spring*
612 *Harbor Laboratory* 2020.07.15.205187 (2020) doi:10.1101/2020.07.15.205187.

613 37. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
614 Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
615 38. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with
616 Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation.
617 *Syst. Biol.* **67**, 216–235 (2018).
618 39. Yang, Z. A space-time process model for the evolution of DNA sequences. *Genetics*
619 **139**, 993–1005 (1995).
620 40. Soubrier, J. *et al.* The influence of rate heterogeneity among sites on the time
621 dependence of molecular rates. *Mol. Biol. Evol.* **29**, 3345–3358 (2012).
622 41. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new
623 software for selection of phylogenetic informative regions from multiple sequence
624 alignments. *BMC Evol. Biol.* **10**, 210 (2010).
625 42. Hanan, B. B. & Tilton, G. R. 60025: relict of primitive lunar crust? *Earth Planet. Sci. Lett.*
626 **84**, 15–21 (1987).
627 43. Barboni, M. *et al.* Early formation of the Moon 4.51 billion years ago. *Sci Adv* **3**,
628 e1602365 (2017).
629 44. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale
630 phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc.*
631 *Lond. B Biol. Sci.* **370**, 20140335 (2015).
632 45. Morel, B., Kozlov, A. M., Stamatakis, A. & Szöllősi, G. J. GeneRax: A Tool for Species-
633 Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene
634 Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **37**, 2763–2774 (2020).
635 46. Bansal, M. S., Kellis, M., Kordi, M. & Kundu, S. RANGER-DTL 2.0: rigorous
636 reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*
637 **34**, 3214–3216 (2018).
638 47. Davín, A. A. *et al.* Gene transfers can date the tree of life. *Nat Ecol Evol* **2**, 904–909
639 (2018).
640 48. Boussau, B. & Scornavacca, C. Reconciling gene trees with species trees. in

Phylogenetics in the Genomic Era (eds. Scornavacca, C., Delsuc, F. & Galtier, N.) 3.2:1–3.2:23 (No commercial publisher| Authors open access book, 2020).

49. Shih, P. M. & Matzke, N. J. Primary endosymbiosis events date to the later Proterozoic 994 with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 996 (2013).
50. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
51. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
52. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).
53. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
54. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* **37**, D233–8 (2009).
55. Rawlings, N. D., Barrett, A. J. & Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **44**, D343–50 (2016).
56. Saier, M. H., Jr, Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34**, D181–6 (2006).
57. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
59. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
60. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence

669 similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011).

670 61. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
671 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

672 62. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
673 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*
674 **14**, 587–589 (2017).

675 63. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**,
676 1586–1591 (2007).

677 64. R Core Team. *R: A language and environment for statistical computing*. https://www.R-
678 project.org/ (2020).

679 65. Wickham, H. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York,
680 USA, (2016).

681

Figure 1

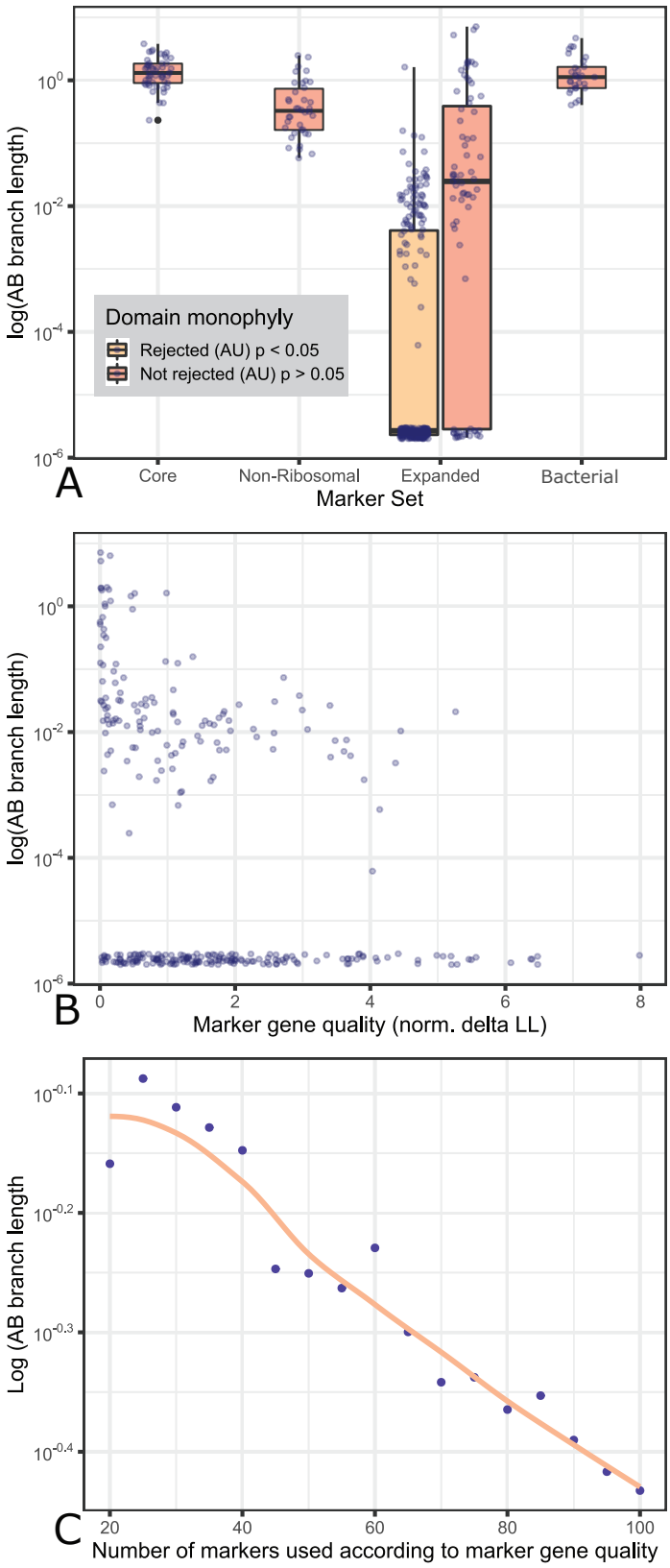


Figure 2

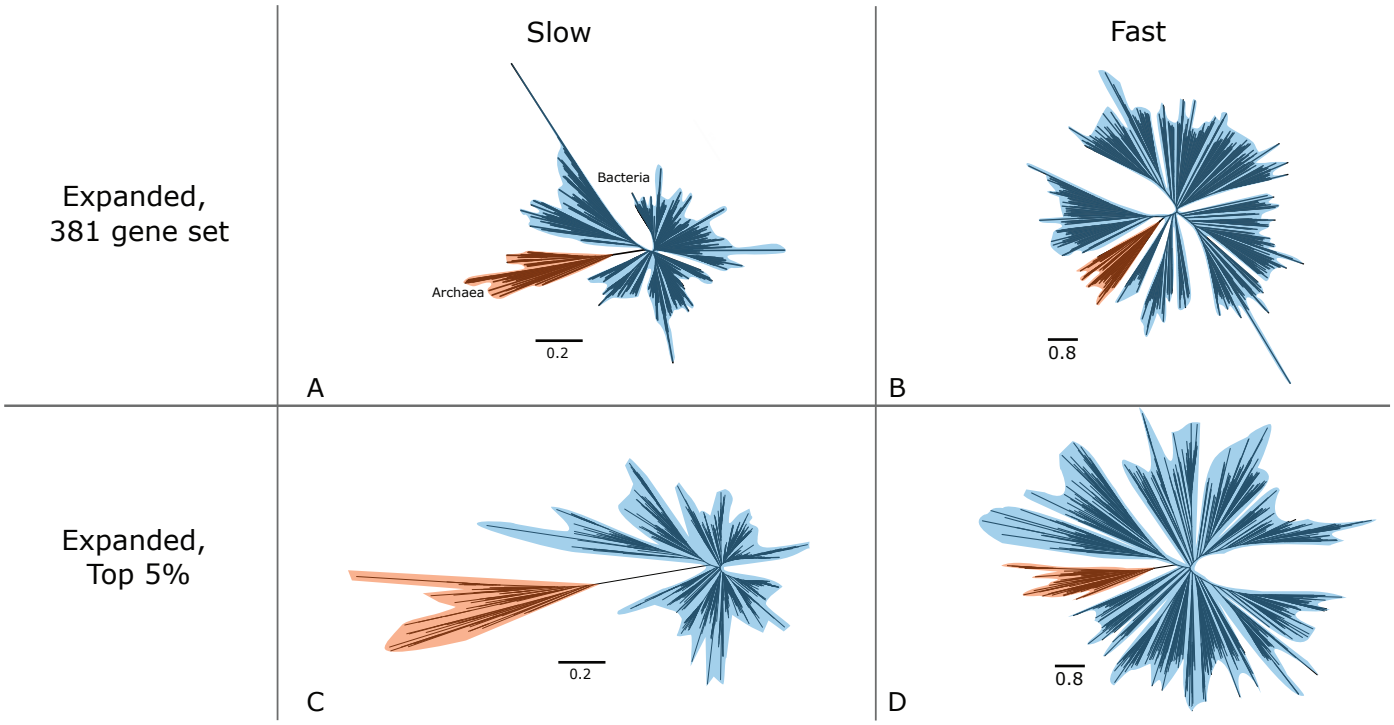
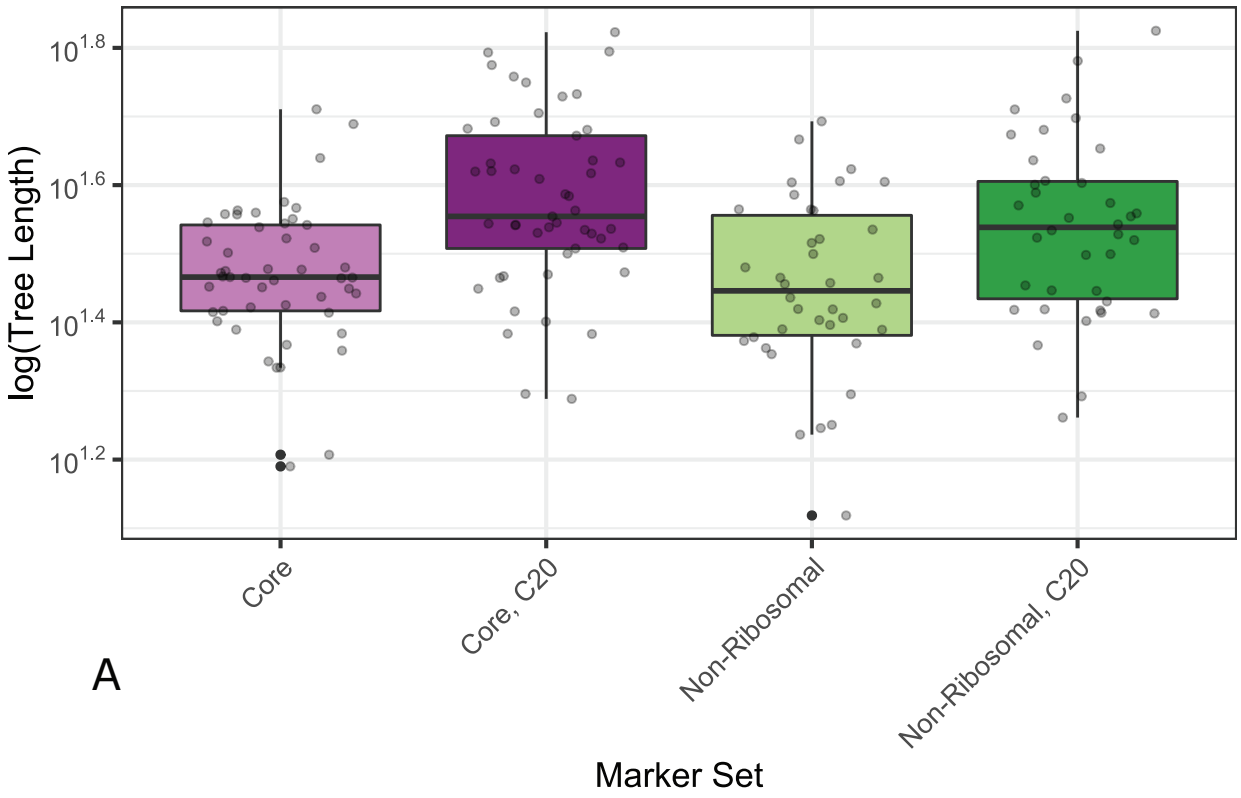


Figure 3



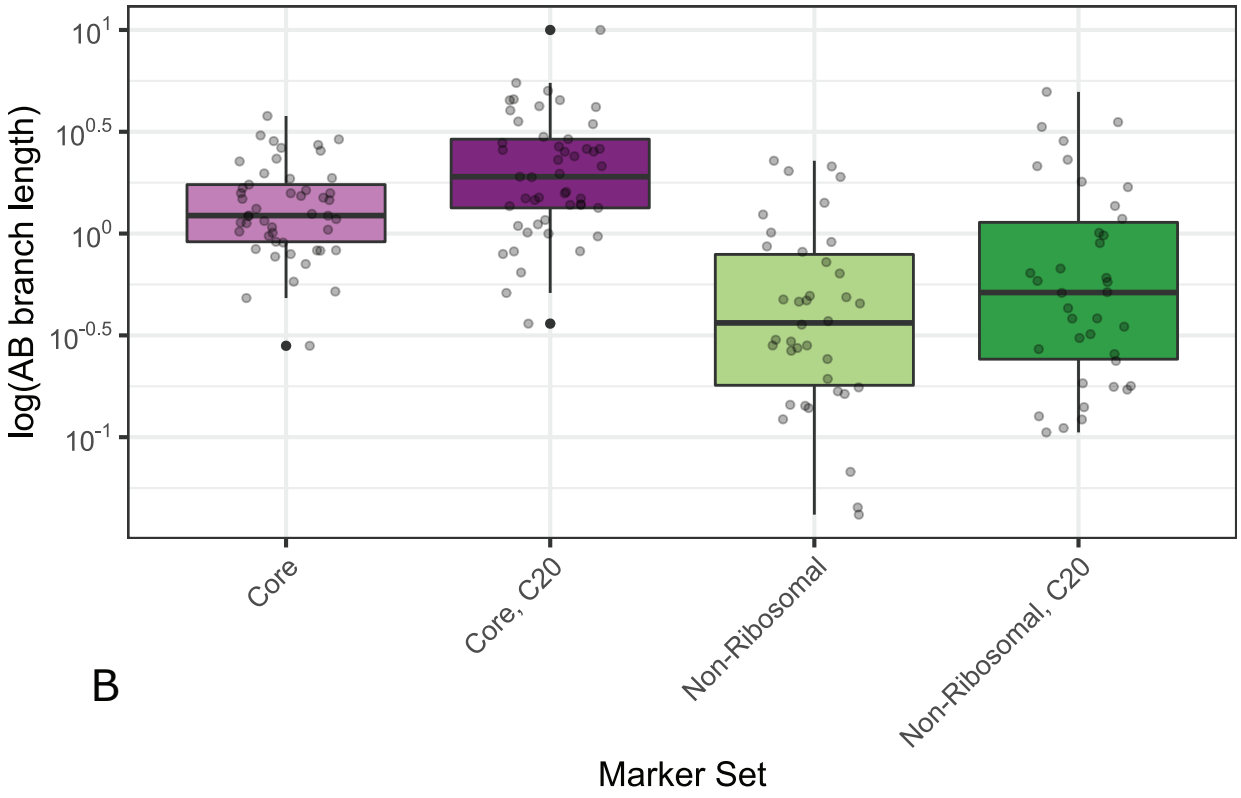


Figure 4

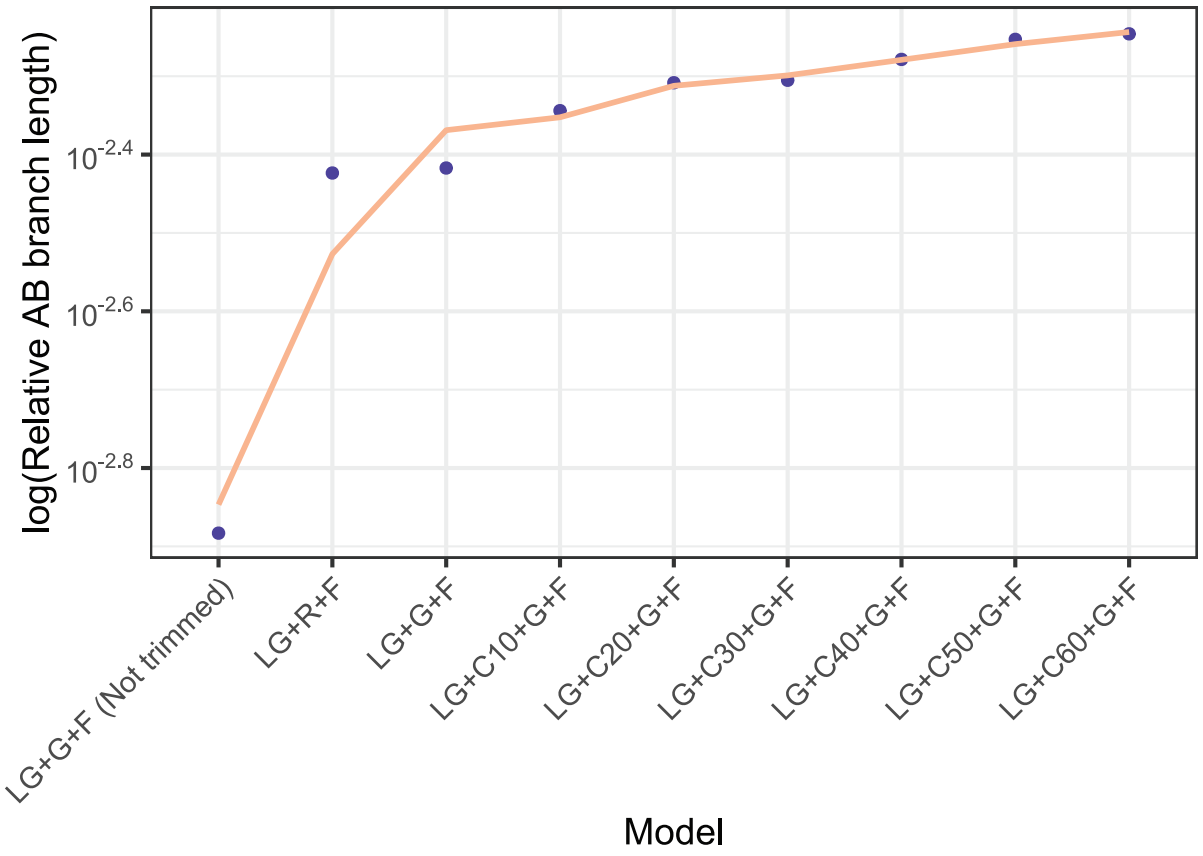


Figure 5

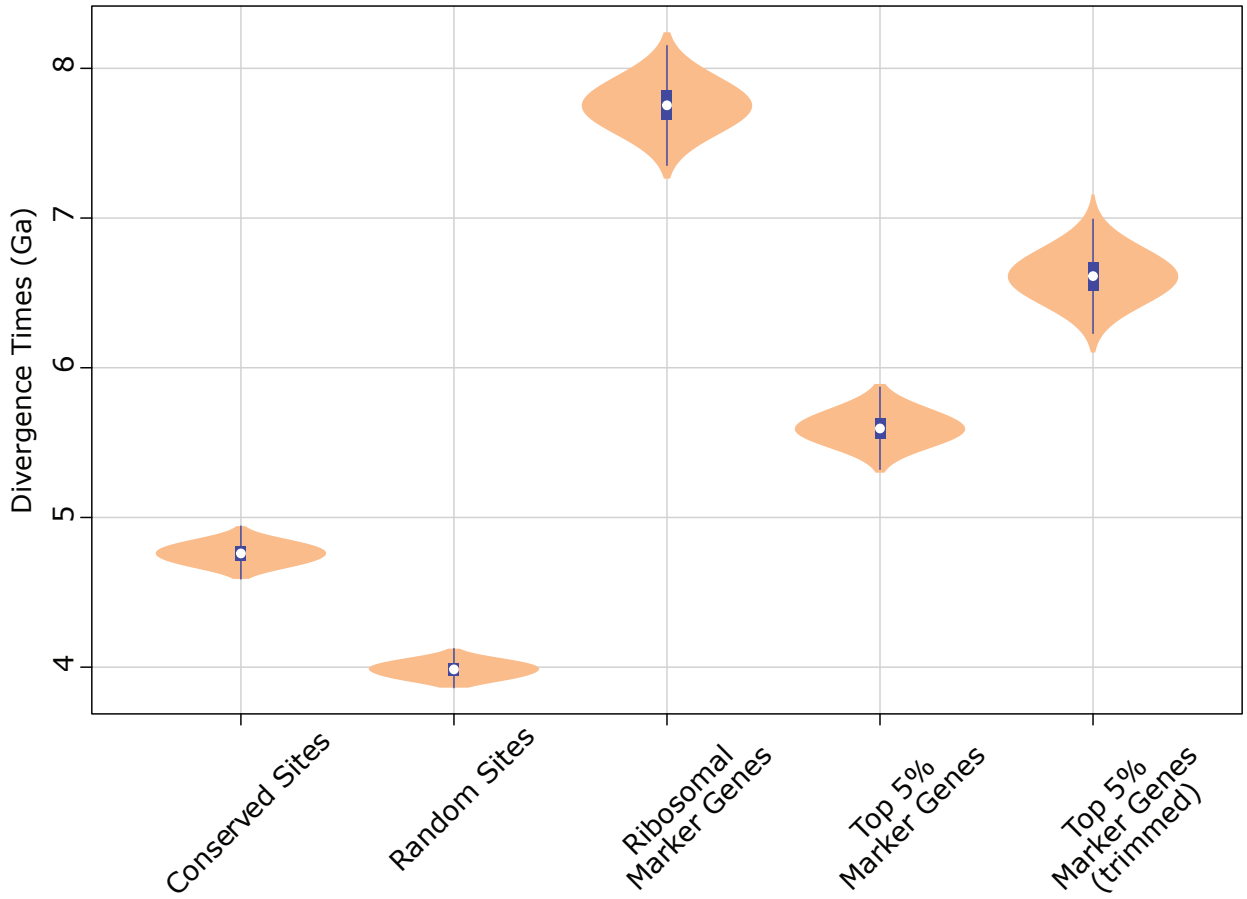


Figure 6