Final version at:
doi/10.1093/molbev/msaa089/5818498

A new analysis of archaea-bacteria domain separation: variable phylogenetic distance and the tempo of early evolution

Sarah J. Berkemer*,[1,5,6] and Shawn E. McGlynn*,[2,3,4]
[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[2]Earth-Life Science Institute, Tokyo Institute of Technology, Meguro, Tokyo, Japan
[3]Blue Marble Space Institute of Science, Seattle, WA, USA
[4]RIKEN Center for Sustainable Resource Science (CSRS)
[5]Bioinformatics Group, Department of Computer Science, University Leipzig, Germany
[6]Competence Center for Scalable Data Services and Solutions Dresden/Leipzig, Germany
***Corresponding author:** E-mail: bsarah@bioinf.uni-leipzig.de, mcglynn@elsi.jp
Associate Editor:

Abstract

Comparative genomics and molecular phylogenetics are foundational for understanding biological evolution. Although many studies have been made with the aim of understanding the genomic contents of early life, uncertainty remains. A study by Weiss *et al.* (2016) identified a number of protein families in the last universal common ancestor of archaea and bacteria (LUCA) which were not found in previous works. Here we report new research that suggests the clustering approaches used in this previous study under-sampled protein families, resulting in incomplete phylogenetic trees which do not reflect protein family evolution. Phylogenetic analysis of protein families which include more sequence homologs rejects a simple LUCA hypothesis based on phylogenetic separation of the bacterial and archaeal domains for a majority of the previously identified LUCA proteins (∼82%). To supplement limitations of phylogenetic inference derived from incompletely populated orthologous groups, and to test the hypothesis of a period of rapid evolution preceding the separation of the domains, we compared phylogenetic distances both within, and between domains, for thousands of orthologous groups. We find a substantial diversity of interdomain vs. intradomain branch lengths, even among protein families which exhibit a single domain separating branch and are thought to be associated with the LUCA. Additionally, phylogenetic trees with long interdomain branches relative to intradomain branches are enriched in information categories of protein families in comparison to those associated with metabolic functions. These results provide a new view of protein family evolution, and temper claims about the phenotype and habitat of the LUCA.

Key words: LUCA, conserved orthologous groups of proteins, orthology, microbial physiology, progenote

Article

Introduction

A longstanding goal of evolutionary biology is to infer the traits of the most ancient

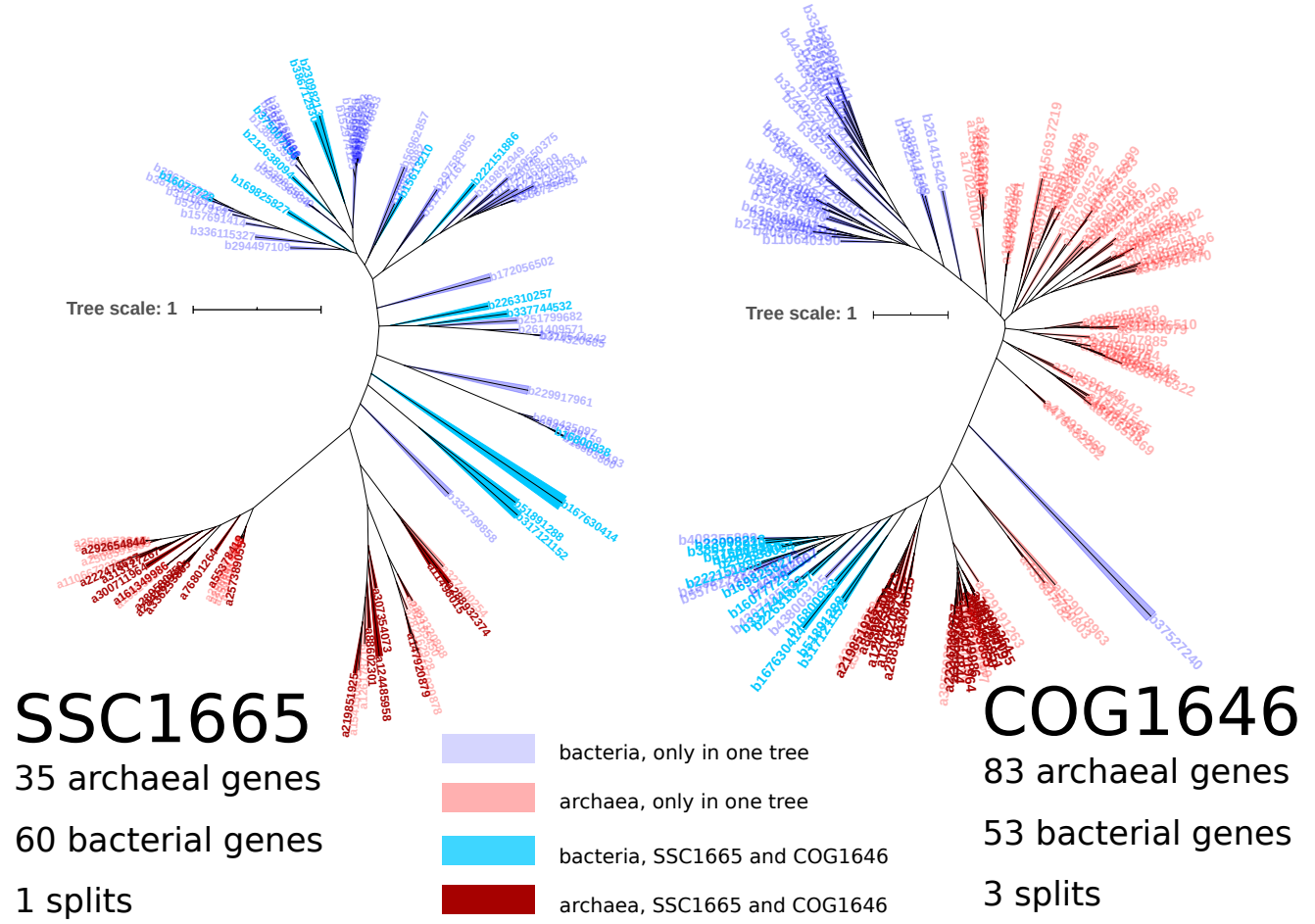

FIG. 1. Comparison of tree toplogies for two trees corresponding to the same protein family, but which contain different collections of sequences (SSC1665 on the left, and COG1646 on the right). Blue colors are bacterial sequences and red colors show archaeal sequences. Sequences with darker color shades appear in both trees, lighter color shaded labels indicates genes that only appear in a single tree. Leaf labels are gene identifiers.

organisms. Conserved presence of a gene in a large number of archaea and bacteria can provide evidence of presence prior to the formation of these two domains, and if phylogenetic analysis indicates domain separation, presence in the LUCA is predicted with greater confidence (Charlebois and Doolittle, 2004; Harris *et al.*, 2003; Koonin, 2003; Woese, 1987; Woese *et al.*, 1990). Although molecular markers such as the 16s ribosomal RNA gene (Woese *et al.*, 1990), ribosomal proteins (Hug *et al.*, 2016), and some nucleotide polymerase subunits such as RpoB (Case *et al.*, 2007) have indicated overall taxonomic relationships upon phylogenetic analysis, comparison of these molecules does not give insight into the metabolisms which power their host cells. To access traits other than those corresponding to these marker genes, gene or protein trees corresponding to metabolic enzymes must be used.

Previous works aimed at identifying protein families associated with the LUCA differ in methodology and conclusions (Becerra *et al.*, 2007; Goldman *et al.*, 2012b). Harris *et al.* (2003) worked with fully sequenced genomes and used the conserved orthologous groups (COGs) (Galperin

et al., 2014; Koonin, 2005; Tatusov *et al.*, 1997) as a protein family reference set for analysis. Their approach was strict, in that they focused on genes present in all complete microbial genomes available at the time; 80 COGs were conserved in the analyzed taxa (Harris *et al.*, 2003) (Table 1). 50 of these conserved COGs separated the archaea, bacteria, and eukaryotic domains upon phylogenetic analysis, suggesting presence in the last common ancestor of those domains. A more recent study (Weiss *et al.* (2016)) involved the analysis of *de novo* clusters of orthologs, and focused on protein families which phylogenetically separated archaeal and bacterial taxa in line with recent data suggesting that eukarya are derived from archaea (Raymann *et al.*, 2015; Zaremba-Niedzwiedzka *et al.*, 2017). There, phylogenetic trees which separated the archaea and bacteria by a single branch were compiled, and broad taxonomic distribution (conservation) was not prioritized in the search for LUCA associated proteins; the presence of an ortholog in two phyla - in addition to phylogenetically separation of the archaea and bacteria - was the set requirement as being a LUCA candidate. Under these criteria, 355 orthologous groups (Single Split Clusters; SSC) were inferred to be present in the common ancestor of archaea and bacteria (Table 1). This latter study was met with some concern (Gogarten and Deamer, 2016). Here we investigated these two previous studies and their contradicting results by re-analyzing original, as well as updated sequence alignments. We also report results from newly developed methods which allow an assessment of interdomain vs. intradomain evolutionary distance to test the hypothesis that ancient protein families may exhibit a long interdomain distance relative to intradomain distances (Catchpole and Forterre, 2019; Forterre, 2006; Woese, 1998).

Results and Discussion

Phylogenetic assessments are sensitive to the number of sequences analyzed

Figure 1 shows two phylogenetic trees corresponding to portions of one protein family, but populated with a different collection of sequences; SSC1665 Weiss *et al.* (2016) corresponds to COG1646 (below we refer to COGs which correspond to SSCs as SSC^{COG}). The SSC shows a single branch (split) separating the archaea and bacteria, but when more sequences are present (as in the COG), three branches separating the domains are observed. As we report below, this loss of archaea:bacteria monophyly in the SSC when more sequences are present is symptomatic of previous work which was used to investigate the protein repertoire of LUCA (Weiss *et al.*, 2016).

Out of the 355 SSCs, 335 families can be assigned to a COG (Weiss *et al.*, 2016). 3 of these corresponding COGS lack archaeal sequences, leaving 332 COGS which correspond to the SSC data set. In 35 SSCs, two or more identified protein families were assigned

Name	Number of protein families	Number of domain separating families	Underlying Data Set
SSC	286514*	355*	clusters created by Weiss *et al.* (2016)
SSC^{COG}	293	52	SSC comprised of corresp. COG sequences
Conserved COGs	80*	50*	COGs, Harris *et al.* (2003)
archaeal and bacterial COGs	2886	661	COGs, Galperin *et al.* (2014)

Table 1. Table listing data sets analyzed in this study. The number of domain separating groups, and the corresponding number of domain separating families found in previous studies are marked by * as reported by Weiss *et al.* (2016) and Harris *et al.* (2003). SSC^{COG} is the set of COGs associated with an SSC; these data and those for the total archaeal and bacterial COGs are based on work reported here. Archaeal and bacterial COGs is the set of COGs which include at least one protein sequence from each domain. For details on the construction of the data sets, see materials and methods (Section) and supplemental Sections 2 and 3.

to the same COG, indicating either that these SSC families are portions of larger protein families, or that the COG contains paralogous sequences (supplemental additional Tables 3 and 4, supplemental Section 2 and supplemental Figure 1). Altogether then, there are 293 unique COGs that can be identified from the original set of 355 SSC. Of these, only 26 protein families are common with the findings of Harris *et al.* (2003) (supplemental Figure 1, supplemental Table 1).

Phylogenetic re-analysis of the same sequence alignments of Weiss *et al.* (2016) suggested instability of branch positions in the previous study, since 40 of the clusters reported to have a single branch separating the archaea and bacteria domains exhibited more than one archaea-bacteria split when trees were constructed with IQTREE (Nguyen *et al.*, 2014) (supporting table 5b) ((the median interdomain branch support value for trees with more than one separating branch was 0.68; the median interdomain branch support value for the original 355 families constructed with IQTree was 0.9). These results from our re-analysis of the same sequence alignments are consistent with a recent report (Catchpole and Forterre, 2019) which did not recover archaea:bacteria monophyly when the sequence alignment of reverse gyrase was re-analyzed. Other studies have also found different results when looking at phylogenetic trees of the same families reported as being in the LUCA Weiss *et al.* (2016). For example the COG of FtsZ was previously highlighted (COG0206) as an example of interdomain horizontal gene transfer (Koonin

and Wolf, 2008), however it is found in the list of domain separating LUCA proteins identified in Weiss *et al.* (2016).

Seeking to understand the origins of these conflicting results, we analyzed the number of sequences obtained with different approaches and found that the SSC alignments contain on average less sequences than the corresponding COGs (Figure 2 (left)). Analyzing phylogenies of COGs which correspond to the SSCs (SSC^{COG}), we found that only 52 trees or $\sim$18% of the SSC which have a unique corresponding COG show a single branch separating the archaeal and bacterial domains (single split topology; $s=1$ and a median branch support at the split nodes of 0.93) (Table 1, Figure 2 (right), supplemental Table 1, additional supplemental Table 5b). The median branch support of branches separating archaea and bacteria in the trees with more than a single split was 0.82, and the majority of trees with very low branch support (<0.4) at domain separating nodes in the SSC were found with increased branch support values in the SSC^{COG}, showing that the addition of orthologs improved branch support for some of the protein families (supplemental figure 15; support values for COG and SSC trees can be found in supplemental Figure 13, the associated tables). These results show that when the small protein families identified earlier (Weiss *et al.*, 2016) are populated with more sequences, the previously reported monophyly between the archaea and bacteria disappears for most of the families. Including COG derived trees which exhibit up to 3 archaea:bacteria branches in their topology, 112 trees (or $\sim$38% of SSC which have a corresponding COG) match with the reported tree topology of archaea bacteria separation reported previously (supplemental Table 1).

In contrast, phylogenetic analysis of the 50 conserved 3-domain split trees obtained in Harris *et al.* (2003) with the most recent COG database reveals that 48 trees show a 2-domain split (Figure 3). This is remarkable, as the study was conducted 18 years ago, and made use of only 34 genomes available at the time. The identified proteins are primarily involved in translation and DNA replication. 16 of the 26 conserved COGs of Harris *et al.* (2003) which overlap with the SSC show a single split between archaea and bacteria upon analysis of the complete set of COGs, whereas 32 conserved COGs which separate the two domains were not identified in the SSC (supplemental Figure 1).

Incorporating phylogenetic domain separation into tree analysis

Obtaining accurate groups of orthologs is challenging (e.g. (Forslund *et al.*, 2017; Galperin *et al.*, 2017)), and as shown above, the analysis of insufficient numbers of sequences can lead to erroneous conclusions. We sought to develop a metric which would aid in overcoming limitations which arise from analyzing incomplete orthologous sets. Long interdomain phylogenetic

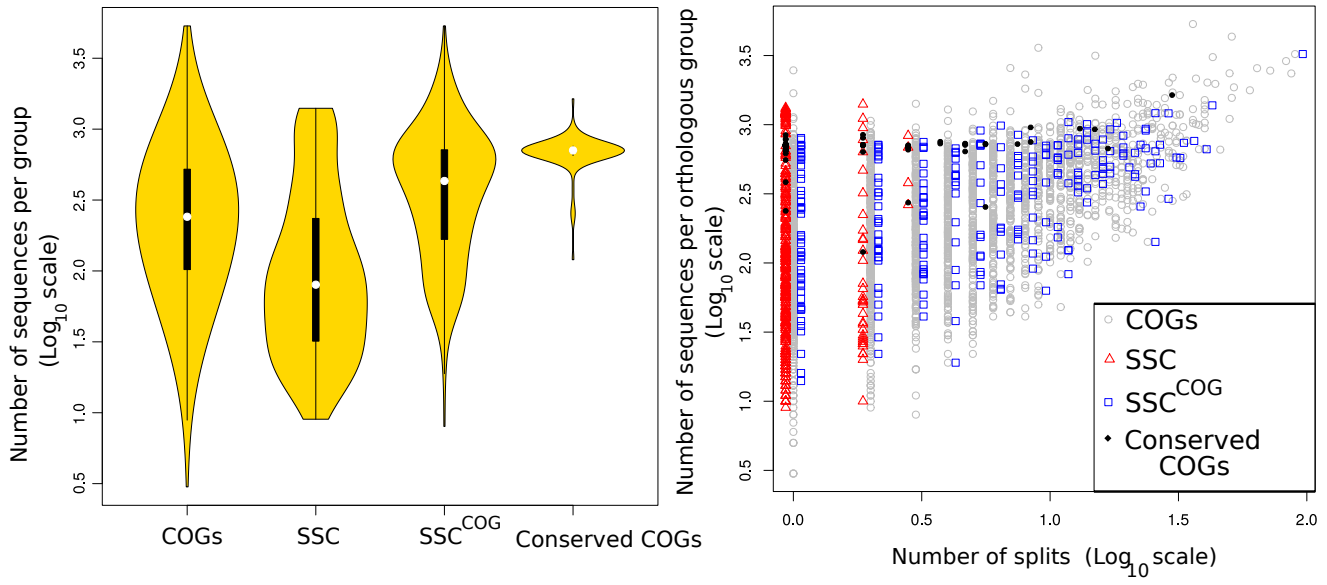

FIG. 2. Left: Violin plot depicting the number of sequences per group discussed in the text and Table 1. **Right**: The number of sequences per orthologous group plotted against the number of interdomain branches (splits) found when the sequences are subjected to phylogenetic analysis (log_{10} scales). The black bar in the yellow area indicates interquartile ranges. Expanding *SSC* (red Squares) with the complete set of sequences of the corresponding COGs results in SSC^{COG} (blue Triangles).

branches may be indicative of a protein family having been in the LUCA, when the tempo of evolution was rapid, whereas families with shorter branches separating the domains may have originated more recently, or exist as examples of recent interdomain gene transfer (Forterre, 2006; Woese, 1998). Under this theory, phylogenetic trees corresponding to protein families present in the archaea and bacteria decedents of the LUCA are predicted to have long inter-domain branches relative to their intra-domain branches. Conversely, protein families which evolved after the separation of the archaea and the bacteria are not predicted to show these long domain separating branches. While this reasoning has previously been applied to a few protein families (Brochier-Armanet and Forterre, 2006; Catchpole and Forterre, 2019), we here developed a quantitative metric and applied it to a large number of protein families.

$\overline{D}$ describes the ratio of intra-domain to inter-domain phylogenetic distances found in a tree (Materials & Methods and supplemental Section 2), and the 3 protein families recently analyzed by Catchpole & Forterre (Catchpole and Forterre, 2019) illustrate the utility of this metric. They analyzed the RNA polymerase beta subunit (RpoB COG0085, $\overline{D}=0.42$), elongation factor G (COG0480, $\overline{D}=0.49$) and reverse gyrase (COG1110, $\overline{D}=0.84$) families and noted the difference in branch lengths separating the domains, suggesting that reverse gyrase is not an ancient protein, whereas RpoB and elongation factor G may be. The $\overline{D}$ value quantifies this previous assessment, although a different sequence set (from the COGs) was used here.

The reverse gyrase COG (COG1110) contains only a portion of the sequences used in the tree reconstructed in Catchpole and Forterre (2019) and shows only two branches separating the archaea and bacteria domains (Figure 3) (Catchpole and Forterre observed 4 interdomain archaea:bacteria branches (splits) with the large alignment). However, the calculated $\overline{D}$ value from the COG is high, suggestive of a more modern protein family which was subject to interdomain gene transfer (Catchpole and Forterre, 2019). Thus, $\overline{D}$ values might be used to supplement phylogenetic inferences based on phylogenetic tree topology, even in the case of incomplete sampling as encountered in this example from the COGs.

Applied to phylogenetic trees drawn from all the COGs, protein families containing a low number of splits between archaea and bacteria groups show variability in $\overline{D}$ values (Figure 3, supplemental Table 1, supplemental additional Tables 3 and 4). Families distributed among archaea and bacteria lineages which display one split and low $\overline{D}$ values include some familiar proteins, for example: ribosomal protein S12 (COG0048, $\overline{D}=$ 0.27, Figure 3), translation elongation factor EF-G (COG0231, $\overline{D}=0.32$), DNA-RNA polymerase RpoB and C (COG0085, $\overline{D}=0.42$ and COG0086, $\overline{D}=0.39$).

Out of COGs which are represented in at least 10 taxa of each domain, 131 of 1751 show a single branch separating archaea and bacteria (supplemental additional Table 5a). Among this list are 63 (∼48%) that are within the information functional categories, including various small ribosomal subunits as listed above (supplemental Figures 6, 7 and 8). Within these protein families exhibiting a single branch separating the archaeal and bacterial domains, variability in $\overline{D}$ exists. Consistent with the finding of variable ages of ribosomal protein components (Kovacs *et al.*, 2017), the ribosomal proteins do not have a coherent $\overline{D}$ value associated between them. For example ribosomal protein S12 (COG0048) appears to be the most domain separating ($\overline{D}=$ 0.27), but ribosomal protein L30/L7a (COG1358), which is known to have nonribosomal function (Cho *et al.*, 2010), shows a $\overline{D}$ value of 0.68. A number of protein families with low $\overline{D}$ values overlap with well separated nearly universal trees (NUTs) (Puigbo *et al.*, 2009), indicating that conservation, phylogenetic domain separation, and long interdomain branches coincide for a set of protein families (supplemental Figure 9 and supplemental additional Table 3).

COGs associated with oxygen metabolism (Liu *et al.*, 2018) all have intra:inter-domain phylogenetic distance ratios $\overline{D}>0.59$, and approach $\overline{D}=1$. (Figure 2, supplemental Figure 10, supplemental Table 2). Surprisingly, $\overline{D}$ was also greater than 0.6 for COGs comprising the four subunits of the CODH/ACS enzyme complex homologous within archaea and bacteria, which in contrast to enzymes involved in oxygen metabolism, are thought to be associated with

the LUCA (Adam *et al.*, 2018; Inoue *et al.*, 2019), or ancient horizontal gene transfers (Inoue *et al.*, 2019). For example, COG2353 (YceI) ($s=1$, $\overline{D}=0.69$) and COG2069 (CdhD) ($s=1$, $\overline{D}=0.61$) (see supplemental Figure 10 and supplemental Table 2 for a full list).

Protein families involved in metabolic processes seem to be less conserved across taxa (Charlebois and Doolittle, 2004), more susceptible to lateral gene transfers (Jain *et al.*, 1999), and do not as frequently display long domain separating branches as those in informational categories, e.g. COG0636 (the Na+ binding c subunits of the ATP synthase ($s=6$, $\overline{D}=0.83$), COG1740 and COG0374 ([Ni-Fe] hydrogenase small and large subunits $s=3$, $\overline{D}=0.78$ and $s=4$, $\overline{D}=0.78$ respectively; see also supplemental Figures 16-19 and supplemental additional Table 3). Some proteins associated with metabolic functions can however be found with lower $\overline{D}$ values (additional Table 3 and 6a); for example the Fe-S oxidoreductase COG1625 ($s=2$, $\overline{D}=0.47$), the beta subunit of Coenzyme F420-reducing hydrogenase COG1035 ($s=1$, $\overline{D}=0.55$) and triosephosphate isomerase COG0149 ($s=2$, $\overline{D}=0.43$) may suggest ancient electron transfer and sugar metabolism. COG1229 (the Formylmethanofuran dehydrogenase subunit A $s=2$, $\overline{D}=0.62$) might also be considered as ancient, but as the number of interdomain split values increase, strong conclusions of the physiology of the LUCA are precluded in the absence of more detailed phylogenetic analysis.

Phylogenetic topology, and domain separation is non random for a set of protein families

The majority of COGs are composed out of ∼10% and ∼90% of genes from each domain (supplemental Figure 11). In a permutation analysis, we took the topology of trees derived from the COGs and shuffled archaea and bacteria in different proportions to create trees of random distributions of archaea and bacteria mapped onto the original trees derived from the COGs. Trees drawn from biological data sets are dramatically different from these random sampling iterations. Only trees which are derived from a low number of sequences (less than 10 genes per group) showed a single split, and the $\overline{D}$ values do not decrease below 0.51 for these single split trees (supplemental Figure 12). For simulated trees with at least 10 genes per group, the minimal number of interdomain branches is 5, which contrasts to the set of COGs where 131 single split trees can be found with at least 10 each of archaea and bacteria (supplemental Section 4, additional supplemental Table 6b and Material & Methods). The observation that domain separating single branches can be obtained by chance in the permutation analysis may be similar to some of the results of Weiss *et al.* (2016), where 184 protein families of the 355 identified have less than 10 sequence representatives of archaea or bacteria (supplemental additional Table 4). False positives

happen by chance more often when there are less sequences (supplemental figure 12).

Diversity of evolutionary mode and history amongst protein families

It has been suggested that proteins present in the LUCA would have a long interdomain phylogenetic branch, reflecting high evolutionary rates before what Woese referred to as crystallization of the domains (Brochier-Armanet and Forterre, 2006; Forterre, 2006; Woese, 1998). This does appear to be reflected in some LUCA protein families (Catchpole and Forterre, 2019), and our analysis of $\overline{D}$ values is a broad test of this hypothesis.

The majority of protein families have intradomain branch lengths that are less than or equal to the interdomain distance, and information categories of proteins are enriched in trees with long interdomain branches. That branch lengths between the archaea and bacteria domains are generally longer than within domain branch lengths is consistent with a hypothesis of a high tempo of evolution prior to the separation of the domains, but the diversity of branch length ratios between protein families is suggestive of unique evolutionary pressures and histories between families. This may be especially relevant considering the diversity of $\overline{D}$ values observed for proteins which are likely to have been in the LUCA (most prominently the ribosomal proteins).

Only a few protein families show $\overline{D}$ values greater than 1. These protein families contain one, or a very small number of sequences from one of the domains (for example COG4694, annotated as the tRNase RloC has only two archaeal sequences, which each result in a archaea:bacteria branch). It could be that these archaeal sequences do not belong in the cluster, or are recent interdomain gene transfers.

While $\overline{D}$ values supplement phylogenetic inference by introducing a distance metric, they do not themselves provide an independent criterion for accessing if a protein family was in the LUCA. It is possible that some proteins may have been in the LUCA but do not show long interdomain branches, and some protein families which were likely in the LUCA simply do not have a single branch separating the domains (Gogarten and Deamer, 2016; Hilario and Gogarten, 1993; Wolf *et al.*, 1999). In many cases simply counting the number of domain separating branches in a phylogenetic tree is insufficient to account for the realities of LGT and loss. Instead, careful phylogenetic analysis is needed to infer protein ancestry, as for example in the case of the CODH/ACS complex (e.g. (Adam *et al.*, 2018; Inoue *et al.*, 2019)).

From the perspective of very early life, it could be that some LUCA proteins might easily undergo interdomain gene transfer, which would blur the ability to recognize them as ancient by a low number of splits. Indeed, Woese's theory of genetic

annealing postulated both mutational rate and lateral gene transfer as components of what may have been a high "temperature" in pre-domain evolution (Woese, 1998). Such easily transferred proteins with "erased" signals of antiquity could be advantageous if early communities relied on horizontal, rather than vertical inheritance (as for example in the stage of a progenote (Woese, 1998; Woese and Fox, 1977)).

Prospectus

Outstanding Orthology Problem

Various approaches exist to detect sets of orthologous sequences, which remains an ongoing challenge (Forslund *et al.*, 2017; Lechner *et al.*, 2014). In our analysis of $\overline{D}$ and the number of interdomain splits, both missing orthologs, and the addition of paralogs in the COGs could affect our results. The COGs are a well known data set *e.g.* (Charlebois and Doolittle, 2004; Goldman *et al.*, 2012a; Harris *et al.*, 2003; Puigbo *et al.*, 2009) created by defining orthology based on sequence comparison and function annotation. This is in line with the orthology conjecture, which states that the most closely related sequences will have the most closely related function (Forslund *et al.*, 2017; Koonin, 2005). Incomplete genome annotation, inaccurate function annotation, and a yet incomplete understanding of the cellular environments where proteins function (Nehrt *et al.*, 2011) make this definition subject for debate. Community efforts to create accurate sets of orthologs (Altenhoff *et al.*, 2016) with increased microbial representation will be critical for future work.

Annotation issues can be corrected by merging bioinformatics with the granularity of biochemistry, but these still confuse analyses aimed at understanding evolution. For example, the putative phosphate acetyltransferase (Pta) sequences found in Weiss *et al.* (2016) lack catalytic residues (Lawrence *et al.*, 2006) and align poorly with the *E. coli* nor *M. thermophila* proteins, meaning that the identified protein is likely different and cannot function as imagined in that report in early energy conservation (*i.e.* conversion).

Concluding Remarks

Our work furnishes a new variable for the assessment of protein family evolution which compliments previous approaches based on conserved presence and phylogenetic topology. Using phylogenetic tree based approaches of the type used here, only limited information can be gained about the LUCA, leaving specific details on physiology largely speculative. Analysis of proteins such as the reverse gyrase, hydrogenase, and nitrogenase discussed here and elsewhere (Boyd *et al.*, 2011a, b; Catchpole and Forterre, 2019) does not support the conclusion of a thermophilic, nitrogen fixing and hydrogen utilizing LUCA (Weiss *et al.*, 2016).

The evolutionary signal of proteins involved in cellular informational processes appears different than those involved in metabolism, and it could

be that the modularity of energy metabolism is in part responsible for an erosion of signal in this latter category. Many of the protein families involved in transcription and protein synthesis do not appear to display inter-domain modularity (consistent with the complexity hypothesis; (Jain *et al.*, 1999)). Their low split values and broad taxonomic distribution are suggestive of their presence in the LUCA, and their small intra:inter-domain phylogenetic distance ratios may reflect high early evolutionary temperatures.

It may be beneficial to integrate protein structure information to better estimate phylogenetic distances. In addition, orthologous groups identified by new methods can be usefully referenced and compared to results from other studies. For example, the nearly universal trees (NUTs) are a set of conserved protein families with variable degrees of domain separation (Puigbo *et al.*, 2010; Puigbo *et al.*, 2009). Going further, employing recent phylogenetic methods such as reconciling gene trees with species trees (Altenhoff and Dessimoz, 2012; Hellmuth, 2017) may aid in overcoming problems associated with limited gene distribution among taxa (Charlebois and Doolittle, 2004), however this is dependent on the availability of reliable species trees. In an effort to integrate molecular data into an Earth history context, geochemical data can give further clues about the environmental conditions on early Earth, allowing for phylogenetic-geochemical calibrations to be made, e.g. (Shih *et al.*, 2017; Wolfe and Fournier, 2018). Altogether, analyses integrating data from multiple dimensions might refine the concept of, and the evolutionary scenario suggested by the statistical tree of life (STOL) (Doolittle and Brunet, 2016; O'Malley and Koonin, 2011; Puigbo *et al.*, 2009).

The physiology of the LUCA remains largely unconstrained. A remaining challenge is to understand the evolutionary distance, and molecular differences between the LUCA and the forms of life which came before it (Cornish-Bowden and Cárdenas, 2017; Gogarten and Deamer, 2016).

Materials and Methods

In order to compare different approaches, we downloaded multiple sequence alignments (MSAs) for COGs[1] (Tatusov *et al.*, 1997) and collected corresponding COGs given in Harris *et al.* (2003) and (Catchpole and Forterre, 2019). The phylogenetic trees and alignments used to obtain conclusions in Weiss *et al.* (2016) were not published in that study, and were instead obtained from author contact on the now defunct pubmedcommons site[2]. After downloading all trees and alignments from the former study, they were subsequently used in our analyses. Corresponding gene families in the COG data set were given for 335 of the 355 clusters identified in (Weiss *et al.*, 2016). We used `FastTree` (Price *et al.*, 2010) with default parameters and `IQTREE`

[1] `https://www.ncbi.nlm.nih.gov/COG`
[2] `ftp://ftp.ncbi.nlm.nih.gov/pubmed/pubmedcommons`

(Nguyen *et al.*, 2014) with `-bb 1000` for bootstrap support and `-m JTT` specifying the evolutionary model, to reconstruct trees based on the MSAs of Weiss *et al.* (2016) and on MSAs for the dataset of all COGs. The analysis presented in the main text is based on `IQTREE` results, but we also employed FastTree separately and obtained similar results of the . A short comparison between `FastTree` and `IQTREE` results can be found in the supplement, Section 3. The study of Weiss *et al.* (2016) used `RaxML` (Stamatakis, 2014) to build phylogenetic trees, however, we obtained almost the same results (Table 1). We only include COGs in the study that contain archaeal as well as bacterial sequences. This is not the case for COG0050 (the current COG set does not contain archaeal sequences), which is contained in the data set by (Harris *et al.*, 2003), as they additionally included eukaryotic sequences. Therefore, we only include 79 gene families from the study by (Harris *et al.*, 2003). Further information on the data sets can be found in supplemental Section 2. In order to obtain one-to-one orthologs sets for the COGs, obvious paralogous sequences from the same species were removed.

After constructing trees for each COG, we calculated the number of archaea:bacteria branches (splits s) needed to separate archaeal (A) and bacterial (B) genes. This was done with a modified version[3] of the Fitch algorithm (Fitch, 1971). Given a tree, we detect nodes in the tree that represent the lowest common ancestor (lca) of a (possibly maximal) set of archaeal or bacterial species. The trees are binary, thus the parent node p of this lca will have two children nodes, each one spanning a subtree of a different domain. In order to calculate support values for split nodes, we take the support value at p. In case of several splits in the tree, we calculate the average support value. Trees were visualized using `iTOL` (Letunic and Bork, 2006, 2016).

Distances between sequences can be calculated by summing up branch lengths on the path between pairs of leaves of the tree. As shown in Figure 3, we assume a tree to show a single split when at least one of the groups is closely connected, thus average pairwise distances are relatively small. Therefore, we calculated the mean phylogenetic pairwise distances between leaves for intra-group genes (between only archaeal or only bacterial sequences) and inter-group distances, that is, the distances between archaeal and bacterial proteins. The following formulas show how calculations were conducted. Here, A is the set of archaeal and B the set of bacterial genes in a tree, with sizes n and m, respectively. The function $d_t(a_i,a_j)$ calculates the distance in the tree t between archaeal genes a_i and a_j and analogously for $d_t(b_i,b_j)$. Then, $\overline{d}_{AA}(t)$ and $\overline{d}_{BB}(t)$ are the mean pairwise distances between archaeal (bacterial) species in tree t.

$$\overline{d}_{AA}(t) = \frac{\sum_{i,j=1}^{n} d_t(a_i,a_j)}{n\cdot(n-1)}, \overline{d}_{BB}(t) = \frac{\sum_{i,j=1}^{m} d_t(b_i,b_j)}{m\cdot(m-1)}$$

[3] `github.com/bsarah/treeSplits`

The same can be done in order to calculate distances between genes from different groups, thus $d_{AB}(t)$ gives the mean pairwise distance between inter-group genes for tree t.

$$\overline{d}_{AB}(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_t(a_i,b_j)}{n \cdot m}$$

For each tree t, there is a set of genes for archaea and a set of genes for bacteria. We calculate the distance between all possible pairs of archaeal gene a and bacterial gene b by summing up over all archaeal and bacterial genes. As the value is dependent on the tree t, we indicate this by writing d_t. These distances can now be used to calculate the ratio of how closely related genes in one group (intra-group) are in comparison to inter-group distances.

$$\overline{D} = \frac{1/2 \cdot (\overline{d}_{AA} + \overline{d}_{BB})}{\overline{d}_{AB}}$$

A further possibility is to only consider the group of genes that has closer mutual relationships replacing the mean value by the minimum:

$$D = \frac{\min(\overline{d}_{AA}, \overline{d}_{BB})}{\overline{d}_{AB}}$$

Values for $\overline{D}$ are at least equal or larger than the corresponding D value. Values for $\overline{D}$ are plotted in Figure 3 and included in further figures and tables in the supplement. Values for D and $\overline{D}$ are also denoted as `Dmin` and `Dav` in the supplemental tables, respectively.

In order to have a randomized reference set of trees, domain identifiers marked at the leaves (A for archaea or B for bacteria) were shuffled on trees built with `FastTree` from the full set of COG alignments which contain archaeal and bacterial sequences. Thus, topology and size were kept and for each tree, we randomly set the labels to A or B. This exercise was performed with three varied proportions of archaea A and bacteria B in the trees: (i) 30% A and 70% B, (ii) 50% A and 50% B, (iii) 90% A and 10% B. For each of the trees in the randomized data sets, the number of splits and values for $\overline{D}$ were calculated. Distribution of values for splits s and $\overline{D}$ compared to COGs are plotted and shown in supplemental Figure 12.

Supplementary Material and Data Availability

Data sets used in this study including reconstructed phylogenetic trees and randomized trees, are available at `www.bioinf.uni-leipzig.de/supplements/19-004`. Supplemental text, and additional supplemental tables are available online at the journal site.

Acknowledgments

S.J.B. was supported by JSPS Summer program/DAAD; S.E.M. acknowledges support by NSF award # 1724300 Collaborative Research: Biochemical, Genetic, Metabolic, and Isotopic Constraints on an Ancient Thiobiosphere, and JSPS KAKENHI Grant Number JP18H01325. We are grateful for comments provided by Peter F. Stadler, Boswell Wing, David Fike, and Grayson Chadwick, and for discussions with Nathaniel Virgo, and Eric Smith. We thank two anonymous reviewers, who's comments improved the manuscript.

References

Adam, P. S., Borrel, G., and Gribaldo, S. 2018. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proceedings of the National Academy of Sciences*, 115(6): E1166–E1173.

Altenhoff, A. M. and Dessimoz, C. 2012. Inferring orthology and paralogy. In *Methods in Molecular Biology*, pages 259–279. Humana Press.

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P., *et al.* 2016. Standardized benchmarking in the quest for orthologs. *Nature methods*, 13(5): 425.

Becerra, A., Delaye, L., Islas, S., and Lazcano, A. 2007. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annual Review of Ecology, Evolution, and Systematics*, 38(1): 361–379.

Boyd, E., Anbar, A., Miller, S., Hamilton, T., Lavin, M., and Peters, J. 2011a. A late methanogen origin for molybdenum-dependent nitrogenase. *Geobiology*, 9(3): 221–232.

Boyd, E. S., Hamilton, T. L., and Peters, J. W. 2011b. An alternative path for the evolution of biological nitrogen fixation. *Frontiers in microbiology*, 2: 205.

Brochier-Armanet, C. and Forterre, P. 2006. Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea*, 2(2): 83–93.

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., and Kjelleberg, S. 2007. Use of 16s rrna and rpob genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.*, 73(1): 278–288.

Catchpole, R. and Forterre, P. 2019. The evolution of reverse gyrase suggests a non-hyperthermophilic last universal common ancestor. *Molecular Biology and Evolution*.

Charlebois, R. L. and Doolittle, W. F. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome research*, 14(12): 2469–2477.

Cho, I.-M., Lai, L. B., Susanti, D., Mukhopadhyay, B., and Gopalan, V. 2010. Ribosomal protein L7ae is a subunit of archaeal RNase P. *Proceedings of the National Academy of Sciences*, 107(33): 14573–14578.

Cornish-Bowden, A. and Cárdenas, M. L. 2017. Life before luca. *Journal of theoretical biology*, 434: 68–74.

Doolittle, W. F. and Brunet, T. D. P. 2016. What is the tree of life? *PLOS Genetics*, 12(4): e1005912.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4): 406–416.

Forslund, K., Pereira, C., Capella-Gutierrez, S., da Silva, A. S., Altenhoff, A., Huerta-Cepas, J., Muffato, M., Patricio, M., Vandepoele, K., Ebersberger, I., Blake, J., Breis, J. T. F., Boeckmann, B., Gabaldón, T., Sonnhammer, E., Dessimoz, C., Lewis, S., Altenhoff, A., Bello, C., Blake, J., Boeckmann, B., Briois, S., Capella-Gutierrez, S., Chalstrey, E., Chiba, H., Conchillo-Solé, O., Daubin, V., DeLuca, T., Dessimoz, C., Dufayard, J.-F., Durand, D., Ebersberger, I., Fernández-Breis, J. T., Forslund, K., Glover, N., Hauser, A., Heller, D., Huerta-Cepas, J., Kaduk, M., Koch, J., Koonin, E. V., Kriventseva, E., Kuraku, S., Lecompte, O., Lespinet, O., Levy, J., Lewis, S., Liebeskind, B., Linard, B., Marcet-Houben, M., Martin, M., McWhite, C., Mekhedov, S., Moretti, S., Muffato, M., Mller, S., Nadia, E.-M., Notredame, C., Patricio, M., Penel, S., Pereira, C., Pilizota, I., Redestig, H., Robinson-Rechavi, M., Schreiber, F., Sjlander, K., Škunca, N., Sonnhammer, E., da Silva, A. S., Steinegger, M., Szklarczyk, D., Thomas, P., Thuer, E., Train, C., Uchiyama, I., Vandepoele, K., Wittwer, L., Xenarios, I., Yates, B., Zdobnov, E., Waterhouse, R. M., and and 2017. Gearing up to handle the mosaic nature of life in the quest for

orthologs. *Bioinformatics*, 34(2): 323–329.

Forterre, P. 2006. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proceedings of the National Academy of Sciences*, 103(10): 3669–3674.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. 2014. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic acids research*, 43(D1): D261–D269.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. 2017. Microbial genome analysis: the COG approach. *Briefings in Bioinformatics*.

Gogarten, J. P. and Deamer, D. 2016. Is luca a thermophilic progenote? *Nature microbiology*, 1(12): 16229.

Goldman, A. D., Baross, J. A., and Samudrala, R. 2012a. The enzymatic and metabolic capabilities of early life. *PLoS One*, 7(9): e39912.

Goldman, A. D., Bernhard, T. M., Dolzhenko, E., and Landweber, L. F. 2012b. Lucapedia: a database for the study of ancient life. *Nucleic acids research*, 41(D1): D1079–D1082.

Harris, J. K., Kelley, S. T., Spiegelman, G. B., and Pace, N. R. 2003. The genetic core of the universal ancestor. *Genome research*, 13(3): 407–412.

Hellmuth, M. 2017. Biologically feasible gene trees, reconciliation maps and informative triples. *Algorithms for Molecular Biology*, 12(1).

Hilario, E. and Gogarten, J. P. 1993. Horizontal transfer of atpase genesthe tree of life becomes a net of life. *Biosystems*, 31(2-3): 111–119.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., *et al.* 2016. A new view of the tree of life. *Nature microbiology*, 1(5): 16048.

Inoue, M., Nakamoto, I., Omae, K., Oguro, T., Ogata, H., Yoshida, T., and Sako, Y. 2019. Structural and phylogenetic diversity of anaerobic carbon-monoxide dehydrogenases. *Frontiers in Microbiology*, 9.

Jain, R., Rivera, M. C., and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7): 3801–3806.

Koonin, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2): 127.

Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39: 309–338.

Koonin, E. V. and Wolf, Y. I. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic acids research*, 36(21): 6688–6719.

Kovacs, N. A., Petrov, A. S., Lanier, K. A., and Williams, L. D. 2017. Frozen in time: The history of proteins. *Molecular Biology and Evolution*, 34(5): 1252–1260.

Lawrence, S. H., Luther, K. B., Schindelin, H., and Ferry, J. G. 2006. Structural and functional studies suggest a catalytic mechanism for the phosphotransacetylase from methanosarcina thermophila. *Journal of Bacteriology*, 188(3): 1143–1154.

Lechner, M., Hernndez Rosales, M., Doerr, D., Wieseke, N., Thvenin, A., Stoye, J., Hartmann, R. K., Prohaska, S. J., and Stadler, P. F. 2014. Orthology detection combining clustering and synteny for very large datasets. *PLoS One*, 9(8): e105015.

Letunic, I. and Bork, P. 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1): 127–128.

Letunic, I. and Bork, P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 44(W1): W242–W245.

Liu, S., Du, M.-Z., Wen, Q.-F., Kang, J., Dong, C., Xiong, L., Huang, J., and Guo, F.-B. 2018. Comprehensive exploration of the enzymes catalysing oxygen-involved reactions and COGs relevant to bacterial oxygen

utilization. *Environmental Microbiology*, 20(10): 3836–3850.

Nehrt, N. L., Clark, W. T., Radivojac, P., and Hahn, M. W. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Computational Biology*, 7(6): e1002073.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1): 268–274.

O'Malley, M. A. and Koonin, E. V. 2011. How stands the tree of life a century and a half after the origin? *Biology Direct*, 6(1): 32.

Price, M. N., Dehal, P. S., and Arkin, A. P. 2010. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3): e9490.

Puigbo, P., Wolf, Y. I., and Koonin, E. V. 2010. The tree and net components of prokaryote evolution. *Genome biology and evolution*, 2: 745–756.

Puigbo, P., Wolf, Y. I., and Koonin, E. V. 2009. Search for a 'tree of life' in the thicket of the phylogenetic forest. *Journal of biology*, 8(6): 59.

Raymann, K., Brochier-Armanet, C., and Gribaldo, S. 2015. The two-domain tree of life is linked to a new root for the archaea. *Proceedings of the National Academy of Sciences*, 112(21): 6670–6675.

Shih, P. M., Ward, L. M., and Fischer, W. W. 2017. Evolution of the 3-hydroxypropionate bicycle and recent transfer of anoxygenic photosynthesis into the chloroflexi. *Proceedings of the National Academy of Sciences*, 114(40): 10749–10754.

Stamatakis, A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312–1313.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. 1997. A genomic perspective on protein families. *Science*, 278(5338): 631–637.

Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W. F. 2016. The physiology and habitat of the last universal common ancestor. *Nature Microbiology*, 1(9): 16116.

Woese, C. 1998. The universal ancestor. *Proceedings of the National Academy of Sciences*, 95(12): 6854–6859.

Woese, C. R. 1987. Bacterial evolution. *Microbiological reviews*, 51(2): 221.

Woese, C. R. and Fox, G. E. 1977. The concept of cellular evolution. *Journal of Molecular Evolution*, 10(1): 1–6.

Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12): 4576–4579.

Wolf, Y., Aravind, L., Grishin, N., and Koonin, E. 1999. Evolution of aminoacyl-trna synthetases analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res*, 9: 689–710.

Wolfe, J. M. and Fournier, G. P. 2018. Horizontal gene transfer constrains the timing of methanogen evolution. *Nature Ecology & Evolution*, 2(5): 897–903.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bckstrm, D., Juzokaite, L., Vancaester, E., Seitz, K. W., Anantharaman, K., Starnawski, P., Kjeldsen, K. U., Stott, M. B., Nunoura, T., Banfield, J. F., Schramm, A., Baker, B. J., Spang, A., and Ettema, T. J. G. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637): 353–358.

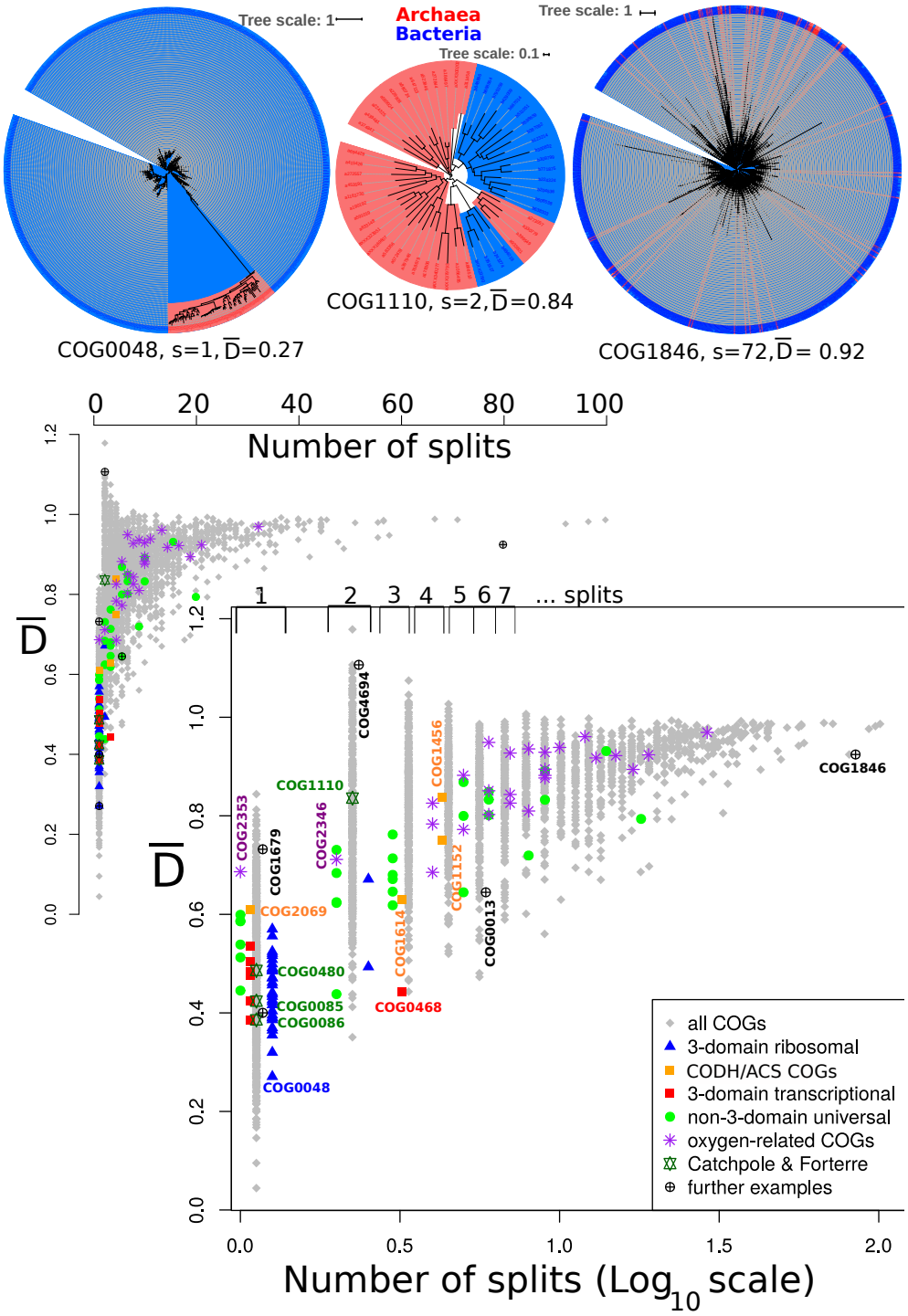

FIG. 3. Relationship between the number of archaea:bacteria interdomain branches (splits) and $\overline{D}$ observed in phylogenetic trees drawn from the COGs. Top: Reconstructed trees for COG0048 (Ribosomal protein S12), COG1110 (Reverse gyrase) and COG1846 (DNA-binding transcriptional regulator, MarR) with corresponding interdomain archaea:bacteria branches (splits) (s) and $\overline{D}$ values. The position of these trees is indicated in part B of the figure. The trees are drawn shading archaea in red color and bacteria in blue color, and the branch lengths are contained within the shaded region.

Bottom: Interdomain split values for each COG plotted against $\overline{D}$, where lower $\overline{D}$ values represent phylogenetic trees with smaller average intra- to inter-domain phylogenetic distances. The inset shows the distribution on normal scale, and the log (split) version is shown below. Symbols are slightly shifted to avoid overlays, and the differently shaped and colored symbols indicate subgroups as defined by Harris *et al.* (2003), Catchpole and Forterre, oxygen related COGs (Liu *et al.*, 2018), CODH/ACS COGs and further examples as indicated in the legend. Brackets on top of the log-plot summarize regions in the plot that correspond to 1, 2,... splits. Labeled symbols refer to corresponding reconstructed phylogenetic trees shown in Figure 3 (top), in supplemental Figure 5, and in additional Table 2. COG0013 is the Alanyl-tRNA synthetase, and COG1679 is a predicted Fe-S cluster binding aconitase