Hindawi Publishing Corporation
Archaea
Volume 2016, Article ID 1851865, 11 pages
http://dx.doi.org/10.1155/2016/1851865

Review Article

Arguments Reinforcing the Three-Domain View of Diversified Cellular Life

Arshan Nasir,[1] Kyung Mo Kim,[2,3] Violette Da Cunha,[4] and Gustavo Caetano-Anollés[5]

[1]*Department of Biosciences, COMSATS Institute of Information Technology, Islamabad, Pakistan*
[2]*Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology, Jeongeup, Republic of Korea*
[3]*Division of Polar Life Sciences, Korea Polar Research Institute, Incheon, Republic of Korea*
[4]*Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de Microbiologie, 75015 Paris, France*
[5]*Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana, IL, USA*

Correspondence should be addressed to Gustavo Caetano-Anollés; gca@illinois.edu

Received 16 August 2016; Revised 18 October 2016; Accepted 3 November 2016

Academic Editor: Stefan Spring

The archaeal ancestor scenario (AAS) for the origin of eukaryotes implies the emergence of a new kind of organism from the fusion of ancestral archaeal and bacterial cells. Equipped with this "chimeric" molecular arsenal, the resulting cell would gradually accumulate unique genes and develop the complex molecular machineries and cellular compartments that are hallmarks of modern eukaryotes. In this regard, proteins related to phagocytosis and cell movement should be present in the archaeal ancestor, thus identifying the recently described candidate archaeal phylum "Lokiarchaeota" as resembling a possible candidate ancestor of eukaryotes. Despite its appeal, AAS seems incompatible with the genomic, molecular, and biochemical differences that exist between Archaea and Eukarya. In particular, the distribution of conserved protein domain structures in the proteomes of cellular organisms and viruses appears hard to reconcile with the AAS. In addition, concerns related to taxon and character sampling, presupposing bacterial outgroups in phylogenies, and nonuniform effects of protein domain structure rearrangement and gain/loss in concatenated alignments of protein sequences cast further doubt on AAS-supporting phylogenies. Here, we evaluate AAS against the traditional "three-domain" world of cellular organisms and propose that the discovery of Lokiarchaeota could be better reconciled under the latter view, especially in light of several additional biological and technical considerations.

1. Introduction

The discovery of the novel candidate archaeal phylum "Lokiarchaeota" from metagenomic samples taken from sites near Loki's Castle hydrothermal vents of the Arctic Ocean was recently reported [1]. There are two interesting aspects to this discovery: (i) several eukaryotic signature proteins (ESPs) related to membrane remodeling, cell division, and the cytoskeleton, previously thought to be either absent or rare in akaryotes (Archaea and Bacteria; *sensu* [2]), were detected in the composite Lokiarchaeota genomes (Loki 1, Loki 2, and Loki 3), and (ii) phylogenomic analyses of concatenated alignment of 36 conserved proteins revealed that eukaryotes and Lokiarchaeota grouped together within Archaea, suggesting an *archaeal ancestor scenario* (AAS) for the origin of eukaryotes [3]. The AAS thus favors a two-domain (2D) view of the tree of life (ToL) where eukaryotes emerge from within Archaea, specifically as sister group to the proposed TACKL (including Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota, and Lokiarchaeota) superphylum [4, 5], after a likely merger of archaeal microbes (resembling Lokiarchaeota) and the mitochondrial ancestors [6].

AAS is fast becoming an accepted scenario to explain deep evolutionary history (e.g., [7–9]) and the origin of eukaryotic cells [10, 11]. Except for some dissenting opinions [12], Lokiarchaeota is now commonly viewed as the "missing link" in the transition from "simple" to "complex" life [1]. However, several key differences in the membrane biology, biochemistry, and virospheres of Archaea and Eukarya seem at odds with AAS (see [13] for a recent review). Simultaneous

ToL reconstructions from concatenated ribosomal proteins and the small-subunit ribosomal RNA (SSU rRNA) gene produced conflicting topologies with the former supporting the AAS while the latter recovering the "Woesian" three-domain (3D) ToL [14] of cellular diversification into domains Archaea, Bacteria, and Eukarya [15]. Because protein sequences are generally more conserved than nucleic acid sequences, SSU rRNA genes possess relatively lower number of informative sites and a higher rate of evolution compared to concatenated ribosomal protein sets. SSU rRNA genes are therefore likely more sensitive to known issues such as the notorious long-branch-attraction (LBA) artifact [16]. In turn, ribosomal proteins exhibit strong compositional biases among the cellular domains of life that need to be better understood [15]. While the study provided an "updated" view of the ToL incorporating hundreds of uncultivated representatives of archaeal and bacterial genera (the so-called "microbial dark matter" [17]) into ToL reconstructions, the authors remained indecisive in picking either the 2D (from concatenated ribosomal proteins) or the 3D (from SSU rRNA) ToL to explain the origin of eukaryotes beyond any doubt [15]. The AAS is also in conflict with several historical phylogenetic and phylogenomic frameworks such as phylogenies built from SSU rRNA sequences [14], single-gene alignments of ancient paralogous genes [18, 19], gene content and order [20, 21], concatenated gene [22] and protein domain [23, 24] sets, and abundance combination and architecture of protein structural domains in modern genomes [23, 25, 26] that have consistently supported the 3D ToL despite disagreements on the location of the root of the ToL and the fact that most generated trees are unrooted [27–30].

It has been argued however that the use of "advanced" models of sequence evolution with relaxed assumptions of homogenous amino acid compositions of gene products across sites and branches is necessary to recover the origin of Eukarya from within Archaea (see [31] for a recent review). However, the presence of distant outgroups (e.g., bacterial ribosomal proteins that are quite divergent from archaeal-eukaryotic counterparts but are used to root the ToLs) and fast-evolving species (e.g., Nanoarchaeota [32] and *Methanopyrus kandleri* [33]) in datasets can make even these sophisticated methods prone to LBA, as shown by recent simulations [29] (see also [34]). Moreover, a concatenated (i.e., supermatrix) approach to phylogenetics, as applied by Spang et al. [1] to support AAS, could be problematic especially when member genes have independent evolutionary histories. Simulations have shown that concatenated gene sets can produce aberrant trees with high bootstrap (BS) support [35]. The approach is also susceptible to heterotachy (i.e., unequal evolutionary rates among genes in a concatenated set) [35, 36], which can complicate inferring deep evolutionary relationships and can introduce distortions to interdomain calculations, among other issues (see Section 5). In light of these considerations, here we examine the evidence supporting the 2D scenario for the diversification of cellular life, perform taxa and character manipulations to reanalyze the dataset of Spang et al. [1] that supported the Lokiarchaeota-Eukarya sisterhood, and consider several biological and technical issues that weaken the 2D in favor of the 3D ToL.

2. Eukaryotic Genomes Are More Complex Than Mere Archaea-Bacteria Genomic Chimeras

AAS remains popular due to the purported chimeric nature of eukaryotic genomes [5, 37]. For example, Guy et al. (2014) wrote, "*The apparent genomic chimerism in eukaryotic genomes is currently best explained by invoking a cellular fusion at the root of the eukaryotes that involves one archaeal and one or more bacterial components*" [3]. Indeed, eukaryotic genomes include many genes that have homologs in Archaea and Bacteria. Genes exhibiting bacterial affinity generally perform metabolic functions while those with archaeal affinity perform informational roles (i.e., DNA replication, transcription, and translation) [37], though exceptions to this "rule" exist (see [38] for a recent review). The proponents of AAS claim that chimerism in eukaryotic genomes is best explained by invoking the transformation of an archaeon (host cell) into a eukaryote by the engulfment of the bacterial ancestor of mitochondria [1]. Thus, a new kind of cell would originate from fusion between two different kinds of cells, a scenario contested to be biologically implausible (see [13] for a recent review).

A coarse-grained examination of eukaryotic genomes also indicates that chimerism is apparently an oversimplification. For example, in addition to Archaea-like and Bacteria-like genes, eukaryotic genomes house a significant number of viral genes and viral-like retrotransposable genetic elements that are likely remnants of ancient viral infections [39, 40]. This viral-like genetic material should therefore imply a "third" partner contributing towards genomic chimerism in eukaryotes. Under AAS, this new partner must invade the eukaryotic genome (or originate de novo) after the proposed fusion event because eukaryotic RNA and retrotranscribing virus families have hitherto not been described in Archaea (see Figure 1 in [41]). This poses a conceptual problem because modern RNA viruses are likely relics of ancient RNA viruses that played significant roles in evolutionary history, perhaps even contributing to the discovery of DNA [42]. Moreover, a substantial number of eukaryotic core genes lack any homologs in akaryotes and were believed to be present in the last common eukaryotic ancestor (up to 40% according to [43]). Remarkably, Eukarya-specific and viral-like genes quantitatively exceed Archaea/Bacteria-like genes in eukaryotic genomes and not all Bacteria-like genes descended from the mitochondrial ancestor (Section 3). At first glance, these observations suggest that the Archaea-Bacteria chimerism is not an a priori requirement to explain eukaryogenesis. Instead, it rather underestimates the distinctive and global nature of eukaryotic genomes.

3. AAS Is Not Supported by Protein Structure Data

A dissection of the proteomic makeup of 383 completely sequenced eukaryal proteomes reveals the global nature of eukaryotic proteomes (Figure 1). A total of 1,661 protein domain fold superfamilies (FSFs) coded by eukaryotic

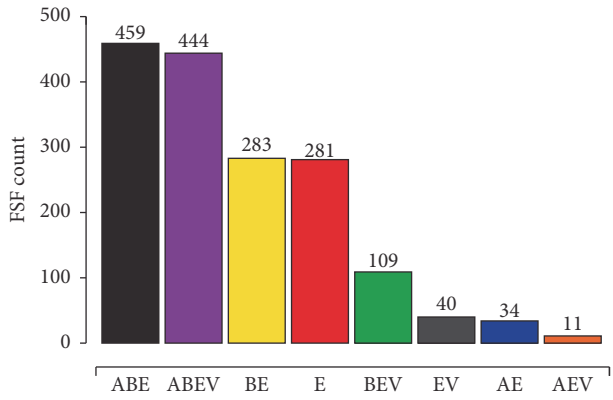

Figure 1: The global nature of eukaryotic proteomes. A total of 1,661 FSFs were detected by the SUPERFAMILY hidden Markov models [44, 45] in proteins coded by 383 completely sequenced proteomes of eukaryotes (FSF assignments of *Lokiarchaeum* were added a posteriori to data taken from [46]). Bars display the number of eukaryotic FSFs that either were shared with Archaea (A) and Bacteria (B) and viruses (V) or were unique to eukaryotes (E).

proteomes can be divided into eight mutually exclusive groups: ABEV (universal), ABE (universal in cells), BEV (all except Archaea), AEV (all except Bacteria), AE (only in Archaea and Eukarya), BE (only in Bacteria and Eukarya), EV (only in Eukarya and viruses), and E (unique to eukaryotes) (Figure 1). FSFs, as defined by the Structural Classification of Proteins (SCOP) database [52, 53], are collections of distantly related protein domains that share recognizable structural and biochemical similarities indicative of divergence from ancestral domain structures. FSFs are thus highly conserved molecular characters that are useful tools to examine deep evolutionary relationships, especially because protein structure is more refractory to change compared to gene and protein sequences that are prone to mutational saturation over long evolutionary distances [54–56].

The AE, BE, and EV groups are of particular interest to this discussion as they imply sharing of homologous FSFs in only two sets of proteomes. The numbers alone are interesting as there is an 8-fold difference in the number of eukaryotic FSFs shared only with Bacteria compared with those shared only with Archaea (283 BE versus 34 AE). This bias challenges both the AAS [1] and the traditionally accepted Archaea/Eukarya sisterhood [14], as one should expect greater sharing between Archaea and Eukarya under these models. Moreover, the EV group even outnumbers the AE FSFs (40 versus 34). While it has been argued that viruses frequently pickpocket cellular genes [57], this historical "belief" has been challenged by several large-scale bioinformatics explorations that suggest gene flow from viruses to cells in fact exceeds gene transfer in the opposite direction [46, 58, 59]. Viruses can also create new genes during intracellular replication using host cell machinery (e.g., ~70–80% of viral genes lack cellular homologs; see Figure 1 in [46]) and some of these genes can later be coopted by cellular genomes (refer to the "virocell" concept [60]). Indeed, 16 out of 38 (42%) EV FSFs perform *Other* functions, a functional category that includes proteins with either unknown or viral functions, suggesting they did not originate in Eukarya (Figure 2). Eukaryotic proteomes also encode a substantial number of unique FSFs (281, ~17% of total eukaryotic FSFs) that confirm that eukaryotic genomes are not mere chimeras of genes mixed from different sources but are more complex than anticipated under the AAS model. In fact, the *Lokiarchaeum* genome (Loki 1) adds only 10 new FSFs to the archaeal repertoire [12] suggesting that the "bridge" between Archaea and Eukarya remains wide, especially when inferring homology at protein structure level.

It can however be argued that the presence of the same FSF in two different sets of proteomes could be due to horizontal gene transfer (HGT) or convergent evolution. However, similar concerns are also applicable to BLAST-based inferences of homology, especially because top BLAST hits are not necessarily orthologous [61]. Importantly, convergent evolution of protein folds is extremely rare [62] because the protein backbone is formed by unique "fingerprint" designs achieved through interactions between amino acid side chains. Due to the direct evolutionary constraint to maintain the overall biochemical function of proteins, disruptions in the protein structural backbone are generally resisted for longer periods of evolutionary time [55, 56, 63]. Moreover, the odds of originating convergent "fingerprints" are very small [62] and there is no reason to suggest that protein structure is relatively more influenced by nonvertical evolution than gene sequences (please see [54] and the references therein). In fact, the recent expansion in the availability of deposited protein structures in structure databases (123,273 structural entries in RCSB Protein Data Bank [64] as of October 5, 2016) offers the unique opportunity to revise life history using an alternative and likely more reliable set of molecular characters.

4. Protein Domain Fold Superfamilies (FSFs) Shared Only by Bacteria and Eukarya (BE) Are Not Restricted to Metabolic Roles

The endosymbiosis of the mitochondrial ancestor likely contributed many metabolic genes to modern eukaryotic genomes [65, 66] and could therefore influence the large size of the BE group (Figure 1). This prompted us to inspect the functional makeup of the AE, BE, and EV groups (Figure 2). Interestingly, BE was not restricted solely to metabolic FSFs but included an ensemble of informational, general, and other FSFs involved in intracellular and extracellular processes (Figure 2). In fact, metabolic FSFs constituted only 31% of BE FSFs (72 out of 233) highlighting the partial contribution of metabolism-inspired gene transfer and enzymatic recruitment to the composition of the BE group. Moreover, eukaryotes shared more informational FSFs with Bacteria than Archaea (29 versus 10). The data therefore suggest that mitochondrial endosymbiosis does not fully account for the large numerical difference in the sizes of BE and AE FSFs. Instead, Bacteria-like eukaryotic genes can alternatively be explained by a combination of (i) endosymbiosis in a protoeukaryotic ancestor (i.e., not an archaeon), (ii) recent HGTs between bacterial and eukaryotic species, and/or (iii)

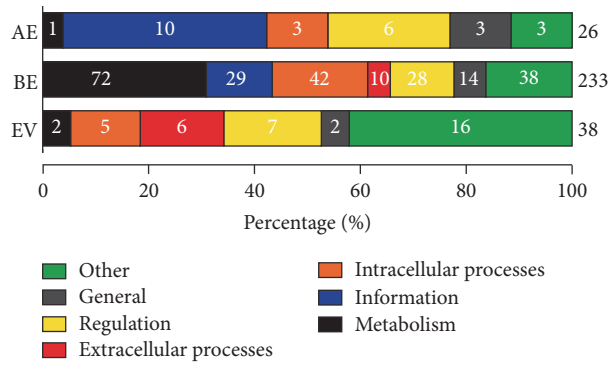

FIGURE 2: Functional composition of AE, BE, and EV Venn groups. FSFs were mapped to one of the seven major functional categories of molecular functions (i.e., *Metabolism*, *Information*, *Regulation*, *Intracellular Processes*, *Extracellular Processes*, *General*, and *Other*), as defined by Christine Vogel (http://supfam.org/SUPERFAMILY/function.html) [47–49]. Numbers on the right indicate total number of FSFs for which functional annotation was available. Numbers on bars indicate total number of FSFs annotated to each of the seven major functional categories in that group.

Bacteria-Eukarya sisterhood in an alternative topology of the 3D ToL [28, 67, 68], without the need to invoke the AAS. It is important to note that, despite several concerns and the use of methods that do not root ToLs (reviewed in [29]), the early origin of Bacteria is taken by default or as a fact under AAS and corresponding phylogenetic trees are rooted using bacterial outgroup sequences. This rooting is ad hoc and could be problematic because it ignores a large body of work challenging the "traditional" bacterial rooting of the ToL [28, 30]. In other words, Bacteria and Eukarya share a wide range of molecular (283 FSFs) and biochemical features (e.g., similar lipid membranes) indicating perhaps a more complex evolutionary history than that explained by chimerism or nonvertical evolution [28].

Similarly, Archaea-like genes in eukaryotes can be explained under the Woesian 3D scenario by invoking a sister group relationship between Archaea and Eukarya, a view historically supported by phylogenies rooted with many paralogous gene sequences [18, 19]. Notably, this topology also accounts for the presence of several ESPs that are scattered in various members of Archaea [1]. Other alternatives involve the origin of the three cellular domains from a complex ancestor of life [69, 70] followed by selective loss of Archaea-like eukaryotic genes in Bacteria and loss of Bacteria-like genes in Archaea (e.g., [71]). For example, the distribution of FSFs in Archaea, Bacteria, Eukarya, and viruses revealed the existence of a shared "universal" core comprising 54% of total FSFs (903 ABE and ABEV FSFs out of a total of 1,661) (Figure 1). The large size of the universal core favors the view that the last common ancestor of cells (and viruses) was already more complex than anticipated (see also [72, 73]). Hence, the differential loss of genes can also account for their absence in one of the three cellular domains of life, especially because many akaryotic species are believed to evolve via genome reduction [74–76]. In summary, even ignoring evidence from FSF distributions, alternative explanations can account for the purported chimerism that is at the root of AAS models suggesting that chimerism could be an oversimplified interpretation of eukaryotic genomes.

5. Technical Issues Related to Taxon and Character Sampling Question AAS

Next, we focus on the more technical aspects of the AAS. It is true that simple genomic comparisons, such as those of FSF distributions, are no substitutes to formal phylogenetic studies (though they have been supported by comparative and phylogenomic exercises [28]). As case study, we evaluated the technical design of the study of Spang et al. [1]. The authors recovered a clade of Lokiarchaeota and Eukarya from trees reconstructed from a concatenated alignment of 36 "universal" proteins in 104 taxa (84 Archaea, 10 Bacteria, and 10 Eukarya, hereafter the 84-10-10 dataset). We focus our discussion on two aspects of their tree reconstruction: (i) taxon sampling and (ii) the use of concatenated alignments (i.e., character sampling and assembly).

Taxon sampling is extremely important for the success of phylogenomic reconstructions as biased and uneven sampling can easily mislead evolutionary interpretations. As Delsuc et al. (2005) wrote, *"garbage in, garbage out"* [77], implying that even the best algorithms can produce false results when taxa/characters do not sufficiently represent extant biodiversity or are known to be problematic. First, overrepresentation of archaeal taxa and sparse selection of bacterial and eukaryal species (i.e., 84-10-10 in [1]) could be problematic, especially because the dataset includes several archaeal species that are sole members of their phylum (e.g., *Candidatus* Korarchaeum cryptofilum), have unknown taxonomic affiliations (e.g., Nanoarchaeota [32, 78]), and/or are fast-evolving (Nanoarchaeota [32], *M. kandleri* [33]). Ideally, taxa should be sampled *randomly*, *equally*, and *densely* from each major group of organisms and increased for reliable tree reconstruction [79, 80] and fast-evolving members excluded [34, 81]. This is showcased by the basal positions of *M. kandleri* and *Thermotoga maritima* within the archaeal and bacterial subtrees in Spang et al.'s (2015) trees (Figure 2 in [1]). *M. kandleri* is a fast-evolving archaeon and its basal position in most phylogenetic trees is now considered a technical artifact [33, 82]. Similarly, the examination of slow-evolving sites in rRNA sequences has revised the phylogenetic placement of *T. maritima* [83] (see also [84]). To dissect these issues, we produced an unrooted distance-based phylogenomic network from the 84-10-10 Archaea-Bacteria-Eukarya concatenated sequence dataset [1]. Interestingly, the network did not group Eukarya within Archaea, recovering instead the 3D view of life (Figure 3(a)). Separately, we reconstructed distance networks from the occurrence (i.e., presence or absence) of universal FSFs (ABE) and FSFs shared by Archaea and Eukarya (34 AE) in 102 taxa sampled *randomly and equally* from the three cellular domains (i.e., 34 taxa each). Again, and despite the AE FSFs biasing reconstructions towards the AAS model, eukaryotes retained their unique identity and did not form a group within the archaeal subtree (Figure 3(b)).

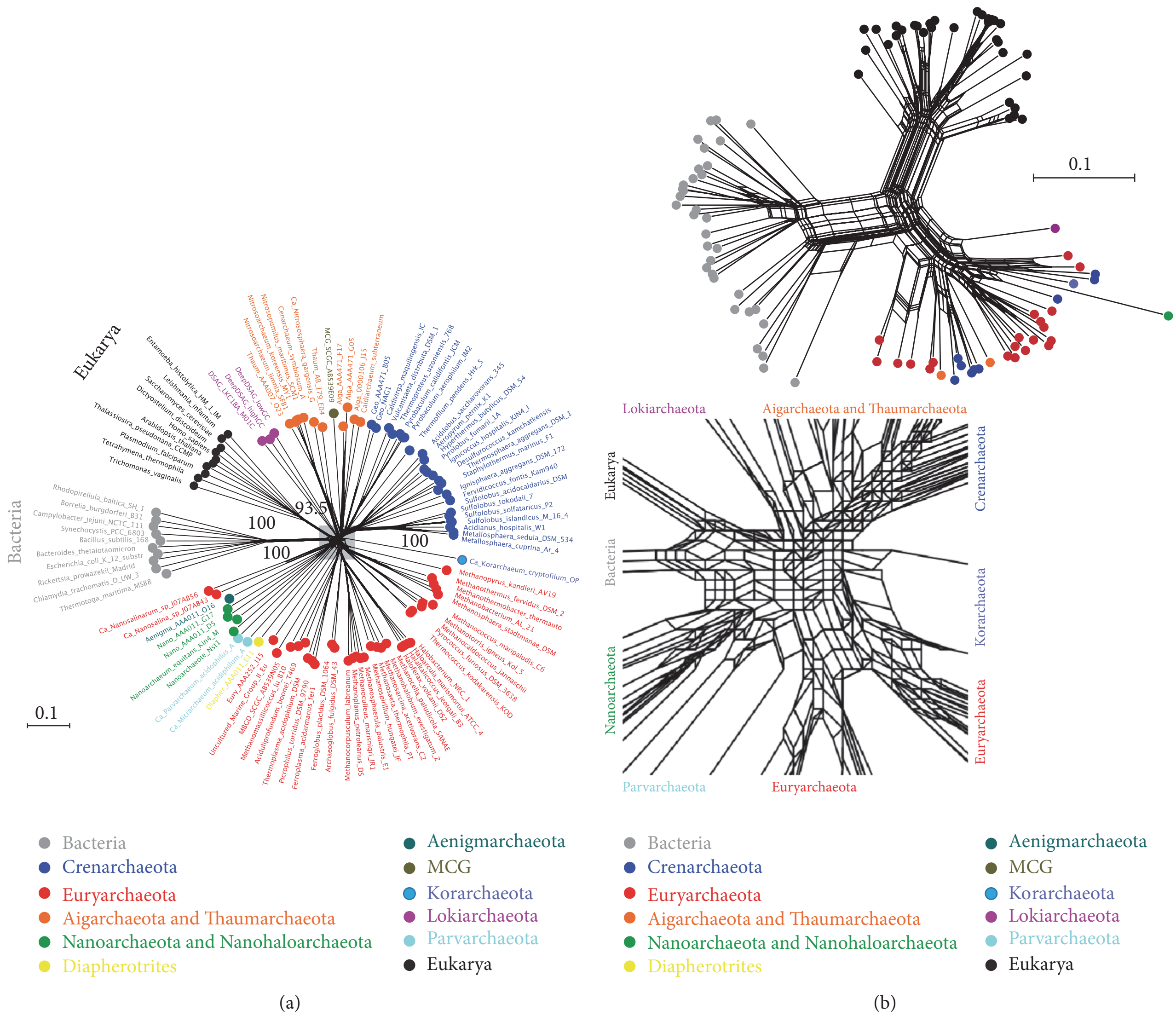

Figure 3: Distance networks do not support AAS. (a) The concatenated alignment of Spang et al. (2015) was provided by Lionel Guy and Thijs Ettema [1]. This alignment concatenated 36 genes (arCOGs) in 104 taxa (84-10-10 dataset) and was already trimmed by authors to remove sites containing >50% gaps. A splits-tree distance-based network (character sites = 10,547, LS fit = 99.97; δ-score = 0.25) reconstructed from the 84-10-10 dataset does not support AAS [1]. Eukaryotic proteomes are in close proximity to Lokiarchaeota but form a monophyletic group of their own. Numbers on branches indicate BS support values for deep split events. The inset shows reticulations at the base of the tree. MCG: miscellaneous crenarchaeotal group. (b) An unrooted splits-tree distance-based network (character sites = 493, LS fit = 99.61; δ-score = 0.24) reconstructed from 34-34-34 dataset sampled from Archaea (including *Lokiarchaeum*), Bacteria, and Eukarya and 493 characters corresponding to presence/absence of FSF domains in the universal ABE (459) and AE (34) groups of Figure 1. For this reconstruction, we only considered organisms exhibiting "free-living" lifestyles since parasitic and obligate parasitic organisms tend to have reduced genomes that are distorted by their holobiont relationship biasing the data matrix. The only "non-free-living" exception was Nanoarchaeota that was added to ensure consistency with the 84-10-10 dataset and to maximize the coverage of archaeal phyla [1]. Both unrooted networks reconstructed by SplitsTree (ver. 4.13.1) [50].

While distance-based methods are no good substitutes to the sophisticated maximum likelihood (ML) and Bayesian analyses (used by Spang et al. [1]) that are less sensitive to LBA and account for relaxed assumptions of amino acid substitutions across sites and branches, they can be useful indicators of underlying conflicts between data and trees and can reveal the existence of reticulations [85]. Importantly, robust retrodictions should provide congruent reconstructions from parametric, nonparametric, and distance methods. Nevertheless, to test the impact of archaeal sampling on the robustness of tree topology, we repeated the phylogenetic analyses by producing 10 new datasets from the 84-10-10 dataset, sampling each time all 10 bacterial and eukaryal species but randomly extracting 10 archaea roughly representative of the known archaeal diversity (i.e., 3 Crenarchaeota, 3 Euryarchaeota, 1 Korarchaeota, 1 Aigarchaeota, 1 Thaumarchaeota, and 1 Lokiarchaeota; Figures S1–S10). *Lokiarchaeum* (Loki 1) was chosen as the Lokiarchaeota representative for these

reconstructions. Despite using the same concatenated alignment of Spang et al. [1], balancing the number of taxa from each domain (i.e., the 10-10-10 datasets) had an immediate effect on the recovered phylogenies. In fact, 7 out of 10 reconstructed ML trees yielded monophyletic Archaea without any mixing of eukaryotic taxa (Figures S2–S8). For the remaining 3 trees that supported paraphyletic Archaea (Figures S1, S9, and S10), we observed that *M. kandleri* (a fast-evolving archaeon) was part of two reconstructions (Figures S9 and S10) indicating that this organism could distort tree topology. For the third tree that recovered paraphyletic Archaea (but in the absence of *M. kandleri*, Figure S1), we observed that group I euryarchaeotes (e.g., Thermococcales and Methanogens group I) were missing among the sampled archaeal taxa. Noticeably, Figure S5 that included *M. kandleri* but did not produce paraphyletic Archaea included both group I (i.e., *Methanococcus maripaludis*, Methanococcales) and group II (*Ferroplasma acidiphilum*, Thermoplasmatales) euryarchaeotes confirming our initial observation that taxon sampling should be broad and inclusive of all groups with careful exclusion of fast-evolving species. Therefore, we produced 3 new phylogenies for the problematic datasets (i.e., Figures S1, S9, and S10) by replacing *M. kandleri* and *Candidatus* K. cryptofilum (the unique member of the putative phylum Korarchaeota, Figures S9 and S10) and *Cand*. K. cryptofilum and *Picrophilus torridus* (a group II euryarchaeote, Figure S1) by two sequences from group I Euryarchaeota (see trees in Figure 4). These revised datasets recovered the monophyly of Archaea (BS > 80%) and produced 3D ToLs (Figure 4). Our experimentation therefore hinted that the AAS (or 2D ToL) could perhaps be an outcome of including fast-evolving species and/or incomplete/unbalanced taxon sampling in phylogenetic datasets that could bias even the latest and sophisticated methods of tree reconstruction. Indeed, recent simulations have revealed that even Bayesian inferences could be prone to LBA when outgroups are too distant [29], a case, for example, when bacterial proteins are used to root ToLs. Indeed, separate ML and Bayesian reconstructions of DNA-dependent RNA polymerase (a universally conserved large protein and a reliable molecular marker [86]) performed after selecting 39 taxa each from Archaea, Bacteria, and Eukarya and after careful exclusion of fast-evolving archaeal species (*Nanoarchaea* and *M. kandleri*) recovered the 3D ToL and a sister relationship between Euryarchaeota and Lokiarchaeota (and its closest evolutionary relative Thorarchaeota [87]) indicating that the result obtained by Spang et al. [1] likely suffered from problematic experimental design (Da Cunha et al. ms. submitted). In summary, both distance-based and probabilistic methods of tree reconstruction and parsimonious inferences drawn from FSF distributions in eukaryotic proteomes challenge the phylogenetic reconstructions of Spang et al. [1] and the AAS model.

The second issue relates to the concatenated or supermatrix approach towards resolving deep evolutionary relationships. Spang et al. (2015) produced a concatenated alignment of 36 conserved genes in 104 taxa. This alignment was trimmed to remove sites with >50% gaps to filter out ambiguous regions. There could be two major problems with this approach: First, trimming using a 50% threshold (partial deletion) is highly dependent on the composition of inclusive taxa. Since the archaeal species dominated the dataset (i.e., 84 out of 104), a minimum of 32 archaeal species must possess the same *indel* present in all bacterial and eukaryal taxa to trim out ambiguous sites. The obvious problem with this approach is that one could trim out different regions when working with different datasets, as these vary in composition of Archaea, Bacteria, and Eukarya. While taxa deletion experiments of Spang et al. [1] claim to minimize the consequences of this issue, balancing the number of organisms sampled from each major group of organisms seems a logical *modus operandi*. Second, concatenated alignments are generally preferred because they yield greater resolution than single-gene markers and are relatively less susceptible to LBA (discussed in [77]). However, their use can be significantly compromised when the genes involved have different evolutionary histories [88], as Spang et al. (2015) themselves noted that the topologies of single-gene phylogenies (which were not shown) were *"often inconclusive with low support values at critical nodes"* [1]. In fact, only 5/36 genes in the concatenated alignment [1] supported the Lokiarchaeota/Eukarya affiliation. Thus, it becomes crucial to reconcile concatenated phylogenies against phylogenies of individual genes (that were included in concatenation) or to perhaps produce alignment-independent phylogenies to avoid these issues [54]. Indeed, several conflicts between concatenated gene sets and single-gene phylogenies specifically aimed towards resolving the phylogenetic relationship between Archaea and Eukarya have historically been reported (reviewed in [13]). To quote Forterre on this topic, "*One should be cautious in the interpretation of trees obtained from the concatenation of protein sequences that produce such contradictory individual trees*" [13]. It can also be a conceptual challenge to visualize the effects of protein domain gain, loss, inversions, and rearrangements in concatenation of several genes. These are well-known evolutionary processes influencing the history of molecular sequences [89] and could pose serious issues especially when primary sequence identity between proteins is very low, as could be the case when comparing distantly related taxa over long evolutionary timespans. Simulations have also shown that concatenated gene sets can lead to inconsistencies and produce misleading trees with high BS values [35], in addition to known issues of heterotachy [36].

Spang et al.'s [1] definition of "universal" proteins is also confusing since some bacterial and eukaryal taxa did not encode one or more of the 36 selected proteins. For example, 7 out of 10 eukaryal taxa did not include the Zn-dependent protease (arCOG04064) [1]. This shows that relatively little phylogenetic information (in terms of both taxa and character sampling) was contributed by bacterial and eukaryal sequences in their study. Moreover, because the dataset included a large number of ribosomal proteins (21 out of 36) that are quite divergent between Bacteria and Archaea/Eukarya, we suspect that the archaeal affiliation of eukaryotes was artificially enhanced under such experimental design (this would be true especially because trees were rooted using bacterial outgroup sequences). Finally, the authors detected several ESPs in the Lokiarchaeota

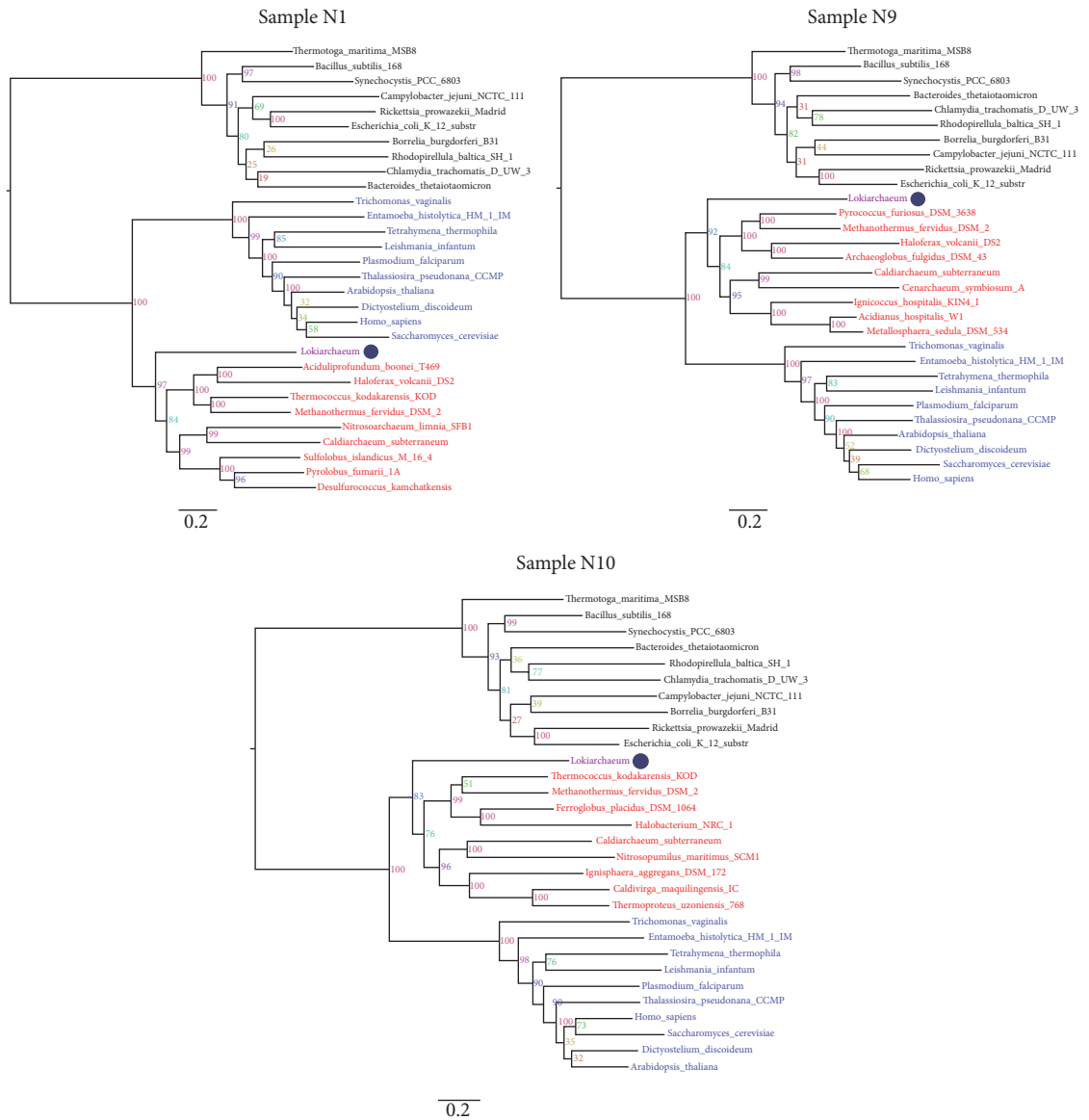

FIGURE 4: ML tree with revised datasets N1, N9, and N10 (see also Figures S1, S9, and S10 in the Supplementary Material available online at http://dx.doi.org/10.1155/2016/1851865). PhyML (ver. 3.1) [51] was used for ML tree reconstruction using LG amino acid substitution model and four categories of evolutionary rates (Γ4). The tree search topology operations were based on the BEST option (both NNI and SPR algorithms). Bacterial, eukaryal, *Lokiarchaeum* (Loki 1), and the rest of the archaeal species are indicated in black, blue, purple, and red, respectively. The purple circle identifies the position of Loki 1. The scale bar represents the average number of substitutions per site. Values at nodes represent support calculated by nonparametric bootstrap (out of 100).

genomes, claiming to be features unique to Lokiarchaeota and Eukarya. However, comparing FSF distributions across the three cellular domains of life and viruses indicates widespread presence of ESPs, especially in viruses (e.g., the Gelsolin-like domain superfamily [12]), suggesting perhaps that archaeal metagenomes were contaminated with eukaryoviruses. The authors also acknowledged the presence of "mimivirus" [90] in the metagenomic sample raising the possibility that its eukaryotic host could also be present. Even if the ESPs genuinely belong to Lokiarchaeota, they can still be explained by the Woesian 3D cellular world by considering a complex archaeal ancestor and subsequent gene loss in modern Archaea [38].

6. AAS Is at Odds with Biochemical and Virosphere Differences between Archaea and Eukarya

To quote Forterre, "*Generally speaking, it is very difficult to resolve ancient relationships by molecular phylogenetic methods for both practical and theoretical reasons, essentially because the informative signal is completely erased at long evolutionary distances*"... "*One possibility to bypass this phylogenetic impasse is to focus on biological plausibility*" [13]. AAS is especially weakened in this regard when one considers differences in the membrane biology and virospheres of Archaea and Eukarya. These issues have been raised before

(e.g., [13, 38, 91–94]) but never satisfactorily addressed by the proponents of AAS. For example, transformation of one kind of cell into another has never been observed in nature even after known cases of HGT across domains (e.g., transfer of about 1,000 genes between Archaea and Bacteria [95]) and endosymbiosis events that are more "intimate" associations between cells but do not produce a new domain of life (e.g., plants remain eukaryotes despite acquiring about one-fifth of their genes from cyanobacteria [96]). Moreover, transformation of an archaeon into a eukaryote would imply transforming archaeal membrane lipids (ether-linked) into bacteria/eukarya like membrane lipids (ester-linked) for which there is no evolutionary rationale. Instead, the difference between the membrane biologies of Archaea and Bacteria/Eukarya could be taken as a powerful synapomorphy supporting the archaeal rooting of the 3D ToL [28]. Moreover, the complex makeup of eukaryotic cells differs greatly from the streamlined makeup of both Bacteria and especially Archaea (please note the substantial number of E FSFs in Figure 1). This gap is only marginally reduced by addition of the *Lokiarchaeum* genome that only adds 10 new FSFs to Archaea [12]. The scenario also seems logically incompatible because of little or no overlap in the genetics and morphology of archaeoviruses and eukaryoviruses (discussed elsewhere [97]). Specifically, many families of RNA viruses that infect eukaryotes seemingly cannot carry out a productive infection cycle in Archaea (though the archaeal virosphere remains largely unexplored [98]). Based on current data, under AAS, one should therefore postulate the late origin of eukaryotic RNA viruses after the transformation had taken place, as claimed by [99]. But this goes against several lines of evidence suggesting that RNA viruses originated very early in evolution and likely led the transition to a DNA world via retrotranscription [42], including a global phylogenomic study of cellular and viral proteomes [46]. The recent discovery of possibly multicellular eukaryotic fossils in 2.1-billion-year-old sediments pushes back in time the last common eukaryotic ancestor [100], further weakening the argument enforcing eukaryotic origins from within Archaea (reviewed in [13]). In short, AAS seems biologically implausible in light of several biological considerations.

7. Conclusions

Metagenomic explorations, development of single-cell sequencing technologies, and improvements in in silico reconstruction of (meta)genomes are yielding novel insights into our understanding of the evolutionary history of cellular organisms. The recent sequencing of Lokiarchaeota composite genomes and resulting phylogenetic analysis suggested an archaeal origin for the eukaryotic cell. The discovery has been widely publicized and the debate surrounding the origin of eukaryotes now considered by many to be settled. However, history inferred from protein structure data reveals a more global picture of the genetic composition of eukaryotic proteomes. Specifically, it takes into account the shared genes with Archaea, Bacteria, and viruses and challenges the purported eukaryotic genomic chimerism that is at the root of AAS models. While some interpret genomic chimerism in eukaryotes by invoking a fusion event at the root of eukaryote evolution, inferences redrawn from phylogenomic analyses performed after balanced taxon and character sampling, removal of fast-evolving species, and comparative analysis of protein structure distribution contradict that interpretation. Moreover, several biological and technical considerations are at odds with the proposed Lokiarchaeota-Eukarya phylogenetic affiliation and suggest that the 3D ToL may still be the more reasonable evolutionary scenario considering biological plausibility and support from molecular data.

Additional Points

The concatenated trimmed alignments for 10-10-10 subsample trees can be downloaded at http://clustomcloud.kopri.re.kr/archaea/Trimmed_alignments_10_10_10.zip.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

Research was supported by grants from the National Science Foundation (OISE-1132791) and the National Institute of Food and Agriculture (ILLU-802-909 and ILLU-483-625) to Gustavo Caetano-Anollés, from the Marine Biotechnology Program (PJT200620, Genome Analysis of Marine Organisms and Development of Functional Applications) funded by the Ministry of Oceans and Fisheries, Korea, to KyungMo Kim, and from the Higher Education Commission, Start-up Research Grant Program (Project no. 21-519/SRGP/R&D/HEC/2014), Pakistan, to Arshan Nasir. Violette Da Cunha is supported by the European Research Council (ERC) grant from the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL-ERC Grant Agreement no. 340440 to Patrick Forterre.

References

[1] A. Spang, J. H. Saw, S. L. Jørgensen et al., "Complex archaea that bridge the gap between prokaryotes and eukaryotes," *Nature*, vol. 521, no. 7551, pp. 173–179, 2015.

[2] P. Forterre, "Neutral terms," *Nature*, vol. 355, no. 6358, p. 305, 1992.

[3] L. Guy, J. H. Saw, and T. J. G. Ettema, "The archaeal legacy of eukaryotes: a phylogenomic perspective," *Cold Spring Harbor Perspectives in Biology*, vol. 6, no. 10, 2014.

[4] L. Guy and T. J. G. Ettema, "The archaeal 'TACK' superphylum and the origin of eukaryotes," *Trends in Microbiology*, vol. 19, no. 12, pp. 580–587, 2011.

[5] J. O. McInerney, M. J. O'Connell, and D. Pisani, "The hybrid nature of the Eukaryota and a consilient view of life on Earth," *Nature Reviews Microbiology*, vol. 12, no. 6, pp. 449–455, 2014.

[6] J. Martijn and T. J. G. Ettema, "From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell," *Biochemical Society Transactions*, vol. 41, no. 1, pp. 451–457, 2013.

[7] T. A. Williams and T. M. Embley, "Changing ideas about eukaryotic origins," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1678, 2015.

[8] T. M. Embley and T. A. Williams, "Evolution: steps on the road to eukaryotes," *Nature*, vol. 521, no. 7551, pp. 169–170, 2015.

[9] E. V. Koonin, "Archaeal ancestors of eukaryotes: not so elusive any more," *BMC Biology*, vol. 13, no. 1, article 84, 2015.

[10] A. Spang and T. J. G. Ettema, "Microbial diversity: the tree of life comes of age," *Nature Microbiology*, vol. 1, no. 5, Article ID 16056, 2016.

[11] E. V. Koonin, "Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier?" *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 370, no. 1678, Article ID 20140333, 2015.

[12] A. Nasir, K. M. Kim, and G. Caetano-Anollés, "Lokiarchaeota: eukaryote-like missing links from microbial dark matter?" *Trends in Microbiology*, vol. 23, no. 8, pp. 448–450, 2015.

[13] P. Forterre, "The universal tree of life: an update," *Frontiers in Microbiology*, vol. 6, article 717, 2015.

[14] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[15] L. A. Hug, B. J. Baker, K. Anantharaman et al., "A new view of the tree of life," *Nature Microbiology*, vol. 1, no. 5, Article ID 16048, 2016.

[16] J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic Zoology*, vol. 27, no. 4, p. 401, 1978.

[17] Y. Marcy, C. Ouverney, E. M. Bik et al., "Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 29, pp. 11889–11894, 2007.

[18] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata, "Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 23, pp. 9355–9359, 1989.

[19] J. P. Gogarten, H. Kibak, P. Dittrich et al., "Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 17, pp. 6661–6665, 1989.

[20] J. O. Korbel, B. Snel, M. A. Huynen, and P. Bork, "SHOT: a web server for the construction of genome phylogenies," *Trends in Genetics*, vol. 18, no. 3, pp. 158–162, 2002.

[21] B. Snel, P. Bork, and M. A. Huynen, "Genome phylogeny based on gene content," *Nature Genetics*, vol. 21, no. 1, pp. 108–110, 1999.

[22] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science*, vol. 311, no. 5765, pp. 1283–1287, 2006.

[23] S. Yang, R. F. Doolittle, and P. E. Bourne, "Phylogeny determined by protein domain content," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 2, pp. 373–378, 2005.

[24] J. Lin and M. Gerstein, "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels," *Genome Research*, vol. 10, no. 6, pp. 808–818, 2000.

[25] M. Wang and G. Caetano-Anollés, "Global phylogeny determined by the combination of protein domains in proteomes," *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2444–2454, 2006.

[26] K. M. Kim and G. Caetano-Anollés, "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms," *BMC Evolutionary Biology*, vol. 12, no. 1, article 13, 2012.

[27] P. Forterre and H. Philippe, "Where is the root of the universal tree of life?" *BioEssays*, vol. 21, no. 10, pp. 871–879, 1999.

[28] G. Caetano-Anollés, A. Nasir, K. Zhou et al., "Archaea: the first domain of diversified life," *Archaea*, vol. 2014, Article ID 590214, 26 pages, 2014.

[29] R. Gouy, D. Baurain, and H. Philippe, "Rooting the tree of life: the phylogenetic jury is still out," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1678, 2015.

[30] H. Brinkmann and H. Philippe, "Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 817–825, 1999.

[31] T. A. Williams, P. G. Foster, C. J. Cox, and T. M. Embley, "An archaeal origin of eukaryotes supports only two primary domains of life," *Nature*, vol. 504, no. 7479, pp. 231–236, 2013.

[32] C. Brochier, S. Gribaldo, Y. Zivanovic, F. Confalonieri, and P. Forterre, "Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?" *Genome Biology*, vol. 6, no. 5, article R42, 2005.

[33] C. Brochier, P. Forterre, and S. Gribaldo, "Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox," *Genome biology*, vol. 5, no. 3, p. R17, 2004.

[34] H. Philippe, H. Brinkmann, D. V. Lavrov et al., "Resolving difficult phylogenetic questions: why more sequences are not enough," *PLoS Biology*, vol. 9, no. 3, Article ID e1000602, 2011.

[35] L. S. Kubatko and J. H. Degnan, "Inconsistency of phylogenetic estimates from concatenated data under coalescence," *Systematic Biology*, vol. 56, no. 1, pp. 17–24, 2007.

[36] B. Kolaczkowski and J. W. Thomton, "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogenous," *Nature*, vol. 431, no. 7011, pp. 980–984, 2004.

[37] M. C. Rivera, R. Jain, J. E. Moore, and J. A. Lake, "Genomic evidence for two functionally distinct gene classes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 11, pp. 6239–6244, 1998.

[38] P. Forterre, "The common ancestor of archaea and eukarya was not an archaeon," *Archaea*, vol. 2013, Article ID 372396, 18 pages, 2013.

[39] A. Katzourakis and R. J. Gifford, "Endogenous viral elements in animal genomes," *PLoS Genetics*, vol. 6, no. 11, Article ID e1001191, 2010.

[40] E. C. Holmes, "The evolution of endogenous viral elements," *Cell Host and Microbe*, vol. 10, no. 4, pp. 368–377, 2011.

[41] A. Nasir, P. Forterre, K. M. Kim, and G. Caetano-Anollés, "The distribution and impact of viral lineages in domains of life," *Frontiers in Microbiology*, vol. 5, article no. 194, 2014.

[42] P. Forterre, "The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells," *Biochimie*, vol. 87, no. 9-10, pp. 793–803, 2005.

[43] L. K. Fritz-Laylin, S. E. Prochnik, M. L. Ginger et al., "The genome of naegleria gruberi illuminates early eukaryotic versatility," *Cell*, vol. 140, no. 5, pp. 631–642, 2010.

[44] J. Gough and C. Chothia, "Superfamily: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments," *Nucleic Acids Research*, vol. 30, no. 1, pp. 268–272, 2002.

[45] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure," *Journal of Molecular Biology*, vol. 313, no. 4, pp. 903–919, 2001.

[46] A. Nasir and G. Caetano-Anollés, "A phylogenomic data-driven exploration of viral origins and evolution," *Science Advances*, vol. 1, no. 8, Article ID e1500527, 2015.

[47] C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S. A. Teichmann, "Supra-domains: evolutionary units larger than single protein domains," *Journal of Molecular Biology*, vol. 336, no. 3, pp. 809–823, 2004.

[48] C. Vogel, S. A. Teichmann, and J. Pereira-Leal, "The relationship between domain duplication and recombination," *Journal of Molecular Biology*, vol. 346, no. 1, pp. 355–365, 2005.

[49] C. Vogel and C. Chothia, "Protein family expansions and biological complexity," *PLoS Computational Biology*, vol. 2, no. 5, article e48, 2006.

[50] D. H. Huson, "SplitsTree: analyzing and visualizing evolutionary data," *Bioinformatics*, vol. 14, no. 1, pp. 68–73, 1998.

[51] S. Guindon, F. Lethiec, P. Duroux, and O. Gascuel, "PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference," *Nucleic Acids Research*, vol. 33, no. 2, pp. W557–W559, 2005.

[52] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Research*, vol. 42, no. 1, pp. D304–D309, 2014.

[53] A. Andreeva, D. Howorth, J.-M. Chandonia et al., "Data growth and its impact on the SCOP database: new developments," *Nucleic Acids Research*, vol. 36, no. 1, pp. D419–D425, 2008.

[54] G. Caetano-Anollés and A. Nasir, "Benefits of using molecular structure and abundance in phylogenomic analysis," *Frontiers in Genetics*, vol. 3, article 172, 2012.

[55] K. Illergård, D. H. Ardell, and A. Elofsson, "Structure is three to ten times more conserved than sequence—a study of structural response in protein cores," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. 3, pp. 499–508, 2009.

[56] D. Lundin, A. M. Poole, B.-M. Sjöberg, and M. Högbom, "Use of structural phylogenetic networks for classification of the ferritin-like superfamily," *Journal of Biological Chemistry*, vol. 287, no. 24, pp. 20565–20575, 2012.

[57] D. Moreira and P. López-García, "Ten reasons to exclude viruses from the tree of life," *Nature Reviews Microbiology*, vol. 7, no. 4, pp. 306–311, 2009.

[58] D. Cortez, P. Forterre, and S. Gribaldo, "A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes," *Genome Biology*, vol. 10, no. 6, article no. R65, 2009.

[59] V. Daubin, E. Lerat, and G. Perrière, "The source of laterally transferred genes in bacterial genomes," *Genome biology*, vol. 4, no. 9, p. R57, 2003.

[60] P. Forterre, "Manipulation of cellular syntheses and the nature of viruses: the virocell concept," *Comptes Rendus Chimie*, vol. 14, no. 4, pp. 392–399, 2011.

[61] L. B. Koski and G. B. Golding, "The closest BLAST hit is often not the nearest neighbor," *Journal of Molecular Evolution*, vol. 52, no. 6, pp. 540–542, 2001.

[62] J. Gough, "Convergent evolution of domain architectures (is rare)," *Bioinformatics*, vol. 21, no. 8, pp. 1464–1471, 2005.

[63] S. Balaji and N. Srinivasan, "Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins," *Protein Engineering*, vol. 14, no. 4, pp. 219–226, 2001.

[64] P. W. Rose, A. Prlić, C. Bi et al., "The RCSB protein data bank: views of structural biology for basic and applied research and education," *Nucleic Acids Research*, vol. 43, no. 1, pp. D345–D356, 2015.

[65] L. Sagan, "On the origin of mitosing cells," *Journal of Theoretical Biology*, vol. 14, no. 3, pp. 225–274, 1967.

[66] R. M. Schwartz and M. O. Dayhoff, "Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts," *Science*, vol. 199, no. 4327, pp. 395–403, 1978.

[67] J. T.-F. Wong, "Emergence of life: from functional RNA selection to natural selection and beyond," *Frontiers in Bioscience—Landmark*, vol. 19, pp. 1117–1150, 2014.

[68] H. Xue, K.-L. Tong, C. Marck, H. Grosjean, and J. T.-F. Wong, "Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life," *Gene*, vol. 310, no. 1-2, pp. 59–66, 2003.

[69] D. Penny and A. Poole, "The nature of the last universal common ancestor," *Current Opinion in Genetics and Development*, vol. 9, no. 6, pp. 672–677, 1999.

[70] D. Penny, L. J. Collins, T. K. Daly, and S. J. Cox, "The relative ages of eukaryotes and akaryotes," *Journal of Molecular Evolution*, vol. 79, no. 5-6, pp. 228–239, 2014.

[71] A. Nasir and G. Caetano-Anollés, "Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification," *Archaea*, vol. 2013, Article ID 648746, 13 pages, 2013.

[72] K. M. Kim and G. Caetano-Anollés, "The proteomic complexity and rise of the primordial ancestor of diversified life," *BMC Evolutionary Biology*, vol. 11, no. 1, article 140, 2011.

[73] K. M. Kim, A. Nasir, and G. Caetano-Anollés, "The importance of using realistic evolutionary models for retrodicting proteomes," *Biochimie*, vol. 99, no. 1, pp. 129–137, 2014.

[74] M. Wang, L. S. Yafremava, D. Caetano-Anollés, J. E. Mittenthal, and G. Caetano-Anollés, "Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world," *Genome Research*, vol. 17, no. 11, pp. 1572–1585, 2007.

[75] M. Wang, C. G. Kurland, and G. Caetano-Anollés, "Reductive evolution of proteomes and protein structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, pp. 11954–11958, 2011.

[76] M. Csurös and I. Miklós, "Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model," *Molecular Biology and Evolution*, vol. 26, no. 9, pp. 2087–2095, 2009.

[77] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[78] E. Waters, M. J. Hohn, I. Ahel et al., "The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 22, pp. 12984–12988, 2003.

[79] D. J. Zwickl and D. M. Hillis, "Increased taxon sampling greatly reduces phylogenetic error," *Systematic Biology*, vol. 51, no. 4, pp. 588–598, 2002.

[80] T. A. Heath, S. M. Hedtke, and D. M. Hillis, "Taxon sampling and the accuracy of phylogenetic analyses," *Journal of Systematics and Evolution*, vol. 46, no. 3, pp. 239–257, 2008.

[81] N. Rodríguez-Ezpeleta, H. Brinkmann, G. Burger et al., "Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans," *Current Biology*, vol. 17, no. 16, pp. 1420–1425, 2007.

[82] C. Petitjean, P. Deschamps, P. López-García, D. Moreira, and C. Brochier-Armanet, "Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life," *Molecular Biology and Evolution*, vol. 32, no. 5, pp. 1242–1254, 2015.

[83] C. Brochier and H. Philippe, "Phylogeny: a non-hyperthermophilic ancestor for bacteria," *Nature*, vol. 417, no. 6886, p. 244, 2002.

[84] O. Zhaxybayeva, K. S. Swithers, P. Lapierre et al., "On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5865–5870, 2009.

[85] D. Bryant and V. Moulton, "Neighbor-net: an agglomerative method for the construction of phylogenetic networks," *Molecular Biology and Evolution*, vol. 21, no. 2, pp. 255–265, 2004.

[86] C. Brochier, P. Forterre, and S. Gribaldo, "An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences," *BMC Evolutionary Biology*, vol. 5, article 36, 2005.

[87] K. W. Seitz, C. S. Lazar, K.-U. Hinrichs, A. P. Teske, and B. J. Baker, "Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction," *ISME Journal*, pp. 1696–1705, 2016.

[88] N. Lartillot, H. Brinkmann, and H. Philippe, "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model," *BMC Evolutionary Biology*, vol. 7, no. 1, article S4, 2007.

[89] A. Nasir, K. M. Kim, and G. Caetano-Anollés, "Global patterns of protein domain gain and loss in superkingdoms," *PLoS Computational Biology*, vol. 10, no. 1, 2014.

[90] B. La Scola, S. Audic, C. Robert et al., "A giant virus in amoebae," *Science*, vol. 299, no. 5615, p. 2033, 2003.

[91] C. R. Woese, "Interpreting the universal phylogenetic tree," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 15, pp. 8392–8396, 2000.

[92] T. Cavalier-Smith, "Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution," *Biology Direct*, vol. 5, no. 1, article 7, 2010.

[93] P. Forterre, "A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong," *Research in Microbiology*, vol. 162, no. 1, pp. 77–91, 2011.

[94] C. de Duve, "The origin of eukaryotes: a reappraisal," *Nature Reviews Genetics*, vol. 8, no. 5, pp. 395–403, 2007.

[95] S. Nelson-Sathi, T. Dagan, G. Landan et al., "Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 50, pp. 20537–20542, 2012.

[96] W. Martin, T. Rujan, E. Richly et al., "Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 19, pp. 12246–12251, 2002.

[97] A. Nasir, F.-J. Sun, K. M. Kim, and G. Caetano-Anollés, "Untangling the origin of viruses and their impact on cellular evolution," *Annals of the New York Academy of Sciences*, vol. 1341, no. 1, pp. 61–74, 2015.

[98] B. Bolduc, D. P. Shaughnessy, Y. I. Wolf, E. V. Koonin, F. F. Roberto, and M. Young, "Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated yellowstone hot springs," *Journal of Virology*, vol. 86, no. 10, pp. 5562–5573, 2012.

[99] E. V. Koonin, V. V. Dolja, and M. Krupovic, "Origins and evolution of viruses of eukaryotes: the ultimate modularity," *Virology*, vol. 479-480, pp. 2–25, 2015.

[100] A. E. Albani, S. Bengtson, D. E. Canfield et al., "Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago," *Nature*, vol. 466, no. 7302, pp. 100–104, 2010.