

OPINION

The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?

Simonetta Gribaldo, Anthony M. Poole, Vincent Daubin, Patrick Forterre and Céline Brochier-Armanet

Abstract | The origin of eukaryotes and their evolutionary relationship with the Archaea is a major biological question and the subject of intense debate. In the context of the classical view of the universal tree of life, the Archaea and the Eukarya have a common ancestor, the nature of which remains undetermined. Alternative views propose instead that the Eukarya evolved directly from a bona fide archaeal lineage. Several recent large-scale phylogenomic studies using an array of approaches are divided in supporting either one or the other scenario, despite analysing largely overlapping data sets of universal genes. We examine the reasons for such a lack of consensus and consider how alternative approaches may enable progress in answering this fascinating and as-yet-unresolved question.

The origin of the Eukarya is one of the most fascinating issues in biology. Eukaryotes possess numerous specific features that can be traced back to the last eukaryotic common ancestor (LECA) (BOX 1), but the timing and mechanisms of their appearance are unclear. Understanding the origin of the ancestral lineage leading to the LECA is a key issue, because it profoundly affects our view of the origin of many eukaryote-specific features, as well as our understanding of fundamental aspects of eukaryote cell biology.

The specific evolutionary relationship between the Archaea and the Eukarya is widely accepted (BOX 1), but there are two main opposing views regarding the nature of this relationship. The first view (referred to here as the ‘three primary domains’ (3D) scenario) proposes that the Eukarya and the Archaea are two distinct sister lineages that are derived from an exclusive common ancestor. This view is reflected in the classical universal tree of life, which delineates three primary domains of life — the Bacteria, the Archaea and the Eukarya¹ (FIG. 1a) — and assumes a root in the branch

leading to the Bacteria, as suggested by ancient couples of paralogous genes^{2,3}. A direct implication of the 3D scenario is that the features shared between the Archaea and the Eukarya are ancestral traits that were inherited from their exclusive common ancestor (FIG. 1a).

By contrast, a second view (referred to here as the ‘two primary domains’ (2D) scenario) states that the Eukarya arose from the association of a bacterium and an archaeon. According to this view, the universal tree of life is composed of only two primary domains — the Archaea and the Bacteria — and the Eukarya represent a secondary domain that arose after the first two (FIG. 1b). Numerous hypotheses to describe the 2D scenario have been proposed^{4–6}. These hypotheses differ in the biological mechanisms proposed for the emergence of the Eukarya and in the nature of the archaeal lineages placed at the origin of the Eukarya. On the basis of small-subunit (16S) ribosomal RNA (rRNA) phylogenies, the Archaea have been historically divided into two major phyla: the Crenarchaeota

(including hyperthermophilic and thermophilic lineages, as well as lineages present in marine environments) and the Euryarchaeota (including lineages found in various environments and harbouring a wide diversity of metabolic abilities, such as methanogenesis)^{7,8}. The eocyte hypothesis^{9,10} suggests that the Eukarya originated from an association between a bacterium and a crenarchaeote. Alternative hypotheses propose instead that the Eukarya emerged from symbiosis between a bacterium and a member of the Euryarchaeota. For example, it has been suggested that the archaeal parent would have been an archaeon similar to members of the genus *Thermoplasma*, a euryarchaeal lineage of wall-less, sulphur-respiring organisms that thrive in acidic environments^{11–14}, whereas the hydrogen hypothesis¹⁵ and the syntrophy hypothesis¹⁶ indicate that the archaeal parent would have been a member of a euryarchaeal lineage that carried out methanogenesis. With respect to the 3D scenario, 2D scenarios make a radically different assumption about the evolutionary relationship between the Eukarya and the Archaea, as they imply that the specific features of eukaryotes evolved from the association between a bacterium and a bona fide archaeon (that is, after the diversification of present-day archaea) (FIG. 1b).

Theoretical arguments for and against the 3D and 2D scenarios have been presented (see, for example, REFS 17–21 and BOX 2) and have led to lively debates. In practice, these opposing views can be tested by molecular phylogeny, but in fact ancient evolutionary relationships are very hard to recover by molecular phylogenies²². First, most of the phylogenetic signal that could provide information for very early speciation events — such as the diversification in the Bacteria, the Archaea and the Eukarya — is very weak, because it has been largely erased by successive substitutions. This may affect the resulting trees, all the more so when they are inferred from single genes such as small-subunit (16S or 18S) rRNA. Second, ancient phylogenies are particularly sensitive to tree reconstruction artefacts that may affect the resulting topology. Accordingly, it has been suggested, for example, that the 3D topology of the tree of life based on universal genes

Box 1 | The evolutionary link between the Eukarya and the Archaea

There are numerous features that are specific for eukaryotes and can be traced back to the last eukaryotic common ancestor (LECA), such as the nucleus, the endomembrane system^{62–64}, the mitochondrion^{65,66}, spliceosomal introns^{67,68}, linear chromosomes with telomeres synthesized by telomerase⁶⁹, meiotic sex⁷⁰, sterol synthesis⁷¹, unique cytoskeletal structures⁷² and the capacity for phagocytosis^{73,74}. However, the timing, mechanisms and order of events leading to these key cellular and genetic features are unclear. Therefore, knowing the origin of the ancestral lineage leading to the LECA is a key issue.

Since their discovery 30 years ago, the Archaea have been shown to have an evolutionary link with the Eukarya (for an exhaustive early review, see REF. 75). In particular, many archaeal components of the systems involved in the transmission and expression of genetic information (translation, replication and transcription; also called 'informational systems') show a higher sequence similarity with their eukaryotic homologues than with their bacterial homologues, and in several cases they are absent in the Bacteria. For example, more than 30 ribosomal proteins are specifically shared between the Archaea and the Eukarya but are not present in the Bacteria⁷⁶, and many archaeal translation factors show similarity to their eukaryotic counterparts⁷⁷. Archaeal RNA polymerases are closely related to their eukaryotic counterparts, in terms of both subunit composition and structure⁷⁸, and use very similar promoters and basal transcription factors for initiation⁷⁹. Finally, the four main activities involved in DNA replication — initiation at the origin, priming of Okazaki fragments, synthesis of the new strands, and DNA unwinding — are carried out by sets of enzymes that are shared between the Archaea and the Eukarya but are not homologous to their bacterial counterparts^{80,81}. A number of discoveries in the Archaea have confirmed the evolutionary link of their informational systems with those of the Eukarya, such as the presence of small nucleolar RNA-related RNAs that modify ribosomal RNA^{82–84}, an exosome (for RNA maturation and degradation) with extensive similarity to its eukaryotic counterpart^{46,85}, and eukaryotic-like histones^{86,87}.

Several 'operational systems' (that is, systems involved in housekeeping functions) also seem to be related in the Archaea and the Eukarya, including membrane-based systems such as vacuolar ATPases³, secretion pathways^{46,88} and a recently identified archaeal cell division system that has homology to the eukaryotic endosomal sorting complex^{89,90}. Furthermore, the Archaea and the Eukarya may harbour an evolutionary link in their metabolic systems, for example between phosphopantetheine adenyltransferase in the coenzyme A biosynthesis pathway⁹¹ and between carbamoylphosphate synthetase in the pyrimidine and arginine biosynthesis pathway⁹².

However, the Archaea have many specific characteristics as well. Notably, their membranes are composed of ether-linked lipids, the glycerol phosphate backbone of which has an opposite stereochemistry (glycerol-1-phosphate) with respect to those found in the Bacteria and the Eukarya (glycerol-3-phosphate) and is also synthesized by an unrelated enzyme⁹³. The Archaea also display unique metabolisms such as methanogenesis⁹⁴, unique enzymes such as specific DNA topoisomerases⁹⁵ and DNA polymerases⁹⁶, and unique cell surface structures^{97,98}.

Finally, the Archaea share several features with the Bacteria, such as the 16S, 23S and 5S ribosomal RNAs, a circular chromosomal architecture (a few exceptions are due to secondary modification⁹⁹) with an operonic gene arrangement, a coupling between transcription and translation¹⁰⁰, and capless translational initiation⁷⁵.

is the result of a tree reconstruction artefact^{10,23–27} that makes the Archaea seem to be a monophyletic group (FIG. 1a), whereas in reality this domain should include the Eukarya (FIG. 1b).

However, the available genomic data from archaea and eukaryotes provide ample material to investigate the question through phylogenomic approaches that combine phylogenetic analyses of many genes. Seven recent large-scale phylogenomic studies have investigated the tree of life, some of which specifically addressed the issue of the relationship between the Archaea and the Eukarya^{23–25,28,29}, and others of which addressed this issue only indirectly^{30,31} (TABLE 1). These studies used different approaches for the collection of gene data sets and different methods of

analysis. Surprisingly, despite the use of very similar gene data sets, results are divided: three studies support the view that the Archaea and the Eukarya are two independent monophyletic lineages (the 3D scenario), whereas four support a relationship between the Eukarya and a particular archaeal lineage (the 2D scenario) but disagree about which archaeal lineage is most closely related to eukaryotes. In this Opinion article we analyse these studies in detail to try to understand the reasons underlying such conflicting results. We conclude that it is premature to label any one of these analyses as definitive. Thus, the deluge of genomic data has not yet yielded a resolution to this debate and has, if anything, intensified it. Although genomic data from diverse species continue to accumulate³², the set of genes used

to construct phylogenetic trees is still very small. We therefore consider how the use of alternative sets of markers may contribute to the quest for the eukaryotic origin.

Large-scale phylogenomic analyses

Support for the 3D scenario. The study by Harris *et al.*³¹ aimed to identify the genetic core of the last universal common ancestor (LUCA) by searching for and analysing universal proteins. The authors identified 80 universal proteins from the 3,100 clusters of orthologous groups (COGs) of the COG database³³. These proteins were analysed individually using maximum-parsimony methods, distance methods and maximum-likelihood methods. Of the 80 proteins analysed, 30 had been affected by horizontal gene transfer (HGT), were present in multiple copies (preventing a clear assessment of orthologous genes) and/or produced trees for which statistical support at nodes was low. Of the remaining proteins, 9 gave trees presenting the Archaea as non-monophyletic, whereas 41 gave trees congruent with the small-subunit rRNA universal tree. Therefore, the majority of the proteins inferred to have been present in the LUCA favoured a 3D topology (FIG. 1a).

The interest of this study lies in using the analysis of a large number of universal genes to move beyond a tree of life generated from small-subunit rRNA only. However, in addition to the limited taxonomic sampling available at the time (TABLE 1), the low statistical support for many of the individual trees obtained in this study underscores the concerns about the reliability of universal trees based on single genes.

A way to overcome the limitations of single-gene phylogenetic analyses is to combine the genes to generate one large data set: after identification of orthologues, the alignments corresponding to each gene are joined to provide a single large (concatenated) alignment and this is analysed using standard phylogenetic methods. This is expected to amplify the weak phylogenetic signal contained in each gene^{34,35}. Ciccarelli *et al.* used such a concatenation approach to reconstruct a universal tree of life to study the evolution of each domain³⁰. They began by extracting universal COG protein families from the complete proteomes of 191 taxa belonging to the Bacteria, the Archaea and the Eukarya. Then they removed genes present in multiple copies (for which assessment of orthology was therefore difficult), those of mitochondrial or chloroplast origin and those bearing signals of HGT, leaving only 31 universal

orthologous proteins for analysis. The concatenation of these markers generated a single large alignment of 8,090 amino acid positions, and the resulting maximum likelihood tree gave strong support for the 3D scenario³⁰.

Key strengths of this analysis are the use of a large taxonomic sample (TABLE 1), an objective methodology to identify HGT and a concatenation approach for tree building. However, some methodological choices warrant caution. For example, the authors used an atypical strategy for assembling their data sets: first, each protein family was aligned independently for the Bacteria, the Archaea and the Eukarya; then these three alignments were automatically trimmed to remove ambiguously aligned positions and were concatenated into domain-specific alignments. The three concatenated data sets were then aligned to each other, leading to a universal data set. The authors used this procedure in an attempt to maximize the number of positions in the global alignment and reduce the number of misaligned residues³⁰. However, it is possible that aligning the sequences from each domain separately might have biased the results toward a 3D topology. Moreover, aligning these three concatenations led to the illegitimate alignment of regions of non-homologous proteins. Such regions represent a substantial part of the alignment (around 1,000 positions; C.B.A. and V.D., unpublished observations), weakening the quality of the final data set. These two factors may raise concerns about the reliability of the resulting tree.

A limitation of concatenation approaches is that they require data sets of orthologues with largely overlapping taxonomic sampling, thereby restricting the number of analysed genes; this is especially true when studying the tree of life, because there are not many universal genes. Yutin *et al.*²⁸ applied a different strategy by looking at more genes and by specifically testing support for the 3D versus 2D scenario. From COG, eukaryotic COG (KOG) and archaeal COG (arCOG) databases^{36–38}, the authors identified 577 protein families with representatives in all three domains of life. Among these, only 136 were considered to have sufficient taxonomic sampling and phylogenetic signal to be used to test the 3D versus 2D topology. The authors analysed support for three different tree topologies: the monophyly of the Crenarchaeota and the Euryarchaeota as support for the 3D scenario, the sistership between the Euryarchaeota and the Eukarya as support for the 2D scenario, and the sistership between the Crenarchaeota and

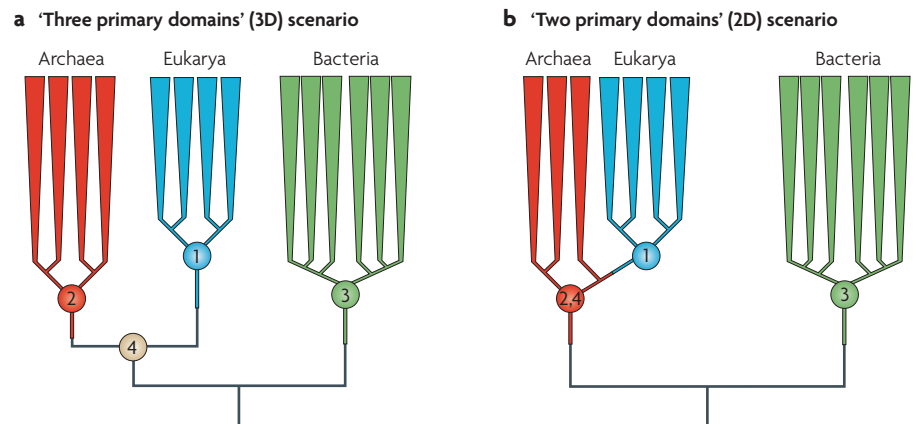


Figure 1 | Relationship between the Eukarya and the Archaea, as inferred from 'three primary domains' and 'two primary domains' scenarios. a | In the 'three primary domains' (3D) scenario, the Eukarya, Archaea and Bacteria form three primary domains, each with a specific most recent common ancestor (1, 2 and 3), and the Eukarya and Archaea are two distinct sister lineages, implying that they share an ancestor (4), the nature of which remains undetermined. The blue bar represents the emergence of specific eukaryotic features. Branch lengths and number of lineages within each domain are arbitrary. For simplicity, the root of the universal tree of life has been placed in the bacterial branch, but note that alternative rootings in the lineage leading to the Eukarya^{102–104} or within the Bacteria^{58,105} have been proposed. This scenario implies that such an ancestor was very likely to be 'prokaryotic' in nature and to harbour all homologous features shared between the Archaea and the Bacteria (for example, it did not have a nucleus). However, this does not mean that this ancestor must have been archaeal in nature (for example, it did not have the typical lipids that define the Archaea today and that probably appeared along the stem leading to this domain). Invoking an archaeal ancestor for eukaryotes would ignore the possibility of evolution along the branch leading to the present-day Archaea. Furthermore, eukaryotic features may have been present in this ancestor and then lost in the branch leading to the Archaea, as some evidence suggests^{76,106,107}. **b** | In the 'two primary domains' (2D) scenario, the Archaea and the Bacteria are the two primary domains, whereas the Eukarya is a secondary domain that arose from the merging of an archaeon and a bacterium (for simplicity, only the relationship between the Archaea and the Eukarya is shown here). In this scenario, the most recent ancestor of the Eukarya (1) derives directly from within the Archaea. Therefore, the most recent ancestor of the Archaea is also the most recent ancestor of the Archaea and the Eukarya (2,4). Note that, similarly to in the 3D scenario, in the 2D scenario there is also a stem along which the characters defining the modern Eukarya (blue bar) would have arisen, similarly to the 3D scenario.

the Eukarya as support for the 2D scenario. In practice, if a given gene strongly supports one particular tree, it should also reject the other two. Unfortunately, only 9 out of the 136 proteins provided clear support for one specific tree and rejected the other two topologies²⁸. Moreover, these nine markers provided a conflicting signal: five supported the 3D scenario, whereas four supported the 2D scenario. Among the remaining 127 markers, 8 rejected only 1 topology and were therefore not able to discriminate between the 2 remaining topologies, whereas 119 rejected none of the topologies and therefore could not discriminate between any of the possible scenarios. These results indicate that the vast majority of the 136 proteins analysed contain insufficient phylogenetic signal to discriminate between any of the 3 possible topologies. Nevertheless, the authors interpreted their data as showing a tendency towards favouring a 3D scenario²⁸.

One strength of this study is the fact that data sets varied in their taxonomic sampling (as compared with the identical taxon sampling that was used in the concatenation approach taken by Ciccarelli *et al.*³⁰), therefore increasing the number of genes that could be analysed. However, the outcome confirms the weakness of the single-gene tree approach: most markers simply do not carry sufficient phylogenetic signal on their own to establish the relationship between the Archaea and the Eukarya.

Support for the 2D scenario. The finding that so little phylogenetic signal is carried in genes across such broad evolutionary distances has encouraged the development of alternative approaches. One of these is to ignore the phylogenetic signal contained in protein or DNA sequences altogether and, instead, use the pattern of presence and absence of homologous genes across different genomes^{39,40}. Rivera & Lake²⁴ investigated

Box 2 | Arguments for 'three primary domains' versus 'two primary domains'

A main argument often presented in favour of the 'two primary domains' (2D) scenarios is that they explain the origin of the Eukarya by referring to known modern archaeal or bacterial species without the need to postulate the existence of now-extinct proto-eukaryotic lineages (in particular, pre-mitochondrial ones) that would have preceded the diversification of present-day eukaryotes, as required by the 'three primary domains' (3D) scenario¹⁷. However, other than the mitochondrion, the complexity of the eukaryote-specific characters that are inferred to have been present in the last eukaryotic common ancestor make it obvious that proto-eukaryotic lineages existed and then became extinct^{20,64}. For there to have been no extinction, all eukaryote-specific features would have had to evolve in the same lineage without any speciation or subsequent diversification occurring, which seems extremely unlikely. Therefore, similarly to 3D scenarios, 2D scenarios imply the existence of a branch (that is, a stem) along which specific eukaryotic features must have evolved²⁰ (see main text, FIG. 1b). The question of the absence of intermediate forms (and therefore of extinct lineages) is not restricted to the Eukarya but applies also to the other two domains, in which proto-archaeal and proto-bacterial lineages preceding the diversification of the present-day Archaea and Bacteria must have existed. A second argument supporting the 2D scenarios is that eukaryotic genomes, in addition to their complement of unique genes, carry a fraction of genes that are more similar to their bacterial counterparts or uniquely shared with the Bacteria and a fraction of genes that are more similar to their archaeal counterparts or uniquely shared with the Archaea, which would testify to a chimaeric origin of the Eukarya. However, it should be noted that 3D scenarios are not incompatible with a substantial genetic contribution from bacteria, as the genes that are more similar to their bacterial counterparts or exclusively shared with the Bacteria could be of mitochondrial origin, acquired through horizontal gene transfer from different bacterial sources or inherited from the last universal common ancestor (LUCA) and then lost or replaced in the Archaea. The eukaryotic genes that are more similar to their archaeal counterparts or shared exclusively with the Archaea might be derived from an exclusive archaeal–eukaryotic common ancestor, acquired through horizontal gene transfer from different archaeal sources or directly inherited from the LUCA but then lost or replaced in the Bacteria. The eukaryote-specific genes might have originated in the eukaryotic lineage or could have been lost or replaced in the Bacteria and the Archaea.

An argument put forward in favour of 3D hypotheses is that, although there is an example of a betaproteobacterium hosting gammaproteobacterial symbionts¹⁰¹, no endosymbiotic associations between an archaeon and a bacterium — a requisite for 2D scenarios — have been described²⁰. A counterargument is that such an association at the origin of the Eukarya would have been a very rare or unique event^{15,16,59,60}. Finally, 2D scenarios must explain the loss or replacement along the eukaryotic stem of typical archaeal features such as their unique lipids as well as specific adaptations such as methanogenesis or hyperthermophily, depending on the archaeal lineage involved in the association⁷³. Possible ways to account for these losses have been proposed^{15,16,59,60}.

the origin of eukaryotes by using an original method of genome content analysis aimed at identifying genome fusions (a method known as 'conditioned reconstruction')⁴¹. The authors applied a probabilistic approach to build trees based on the presence and absence of homologous genes in two bacterial genomes, three archaeal genomes and two eukaryotic genomes (TABLE 1). Their analysis yielded a set of possible trees, each associated with a probability of generating the data. They obtained a most probable tree with a moderate support of 60% and four alternative (less probable) trees, for which support ranged from 1.8% to 16.8%²⁵. The authors reconciled these five conflicting trees by combining them into a cycle graph (a 'ring of life'), which was interpreted to indicate that the Eukarya derive from a genome fusion event between a proteobacterium and a crenarchaeon, therefore providing support for the 2D eocyte hypothesis²⁵.

An interesting feature of this study is that, unlike the other three studies discussed above, it attempts to analyse the origin of eukaryotes by using complete genome data. However, such an approach is not without problems. First, it could be argued that the taxonomic sampling used was insufficient, in particular because it was not representative of the diversity of the domains (the authors used only two yeasts as members of the Eukarya, only one gammaproteobacterium and one firmicute as members of the Bacteria, and only one crenarchaeote and two euryarchaeotes as members of the Archaea; TABLE 1). Second, phylogenetic approaches based on gene presence and absence are known to be highly prone to biases introduced by processes that shape genome contents (such as HGT and gene losses)³⁵. The trees that were used to build the ring may therefore be strongly biased by the numerous HGT events that have occurred between the two

analysed euryarchaeotes and the Bacteria^{42,43}. This may explain why the tree with the highest support showed the gammaproteobacterium and the yeasts to be nested within the Archaea, a relationship which is not consistent with a scenario of eukaryogenesis from a genome fusion between an archaeon and a bacterium²⁴.

Another way to combine the phylogenetic signals carried by different phylogenetic makers is with supertrees. To produce supertrees, the topological information from gene trees, rather than gene sequences, is combined to obtain a composite phylogeny. One popular approach is matrix representation with parsimony (MRP), a method that recodes the branches of gene trees into a matrix of binary characters (1 or 0; for example, each branch defining two groups of sequences, one of which is given the character 1 and the other of which is given the character 0). This matrix is then used to infer the supertree using a parsimony criterion^{34,35,44}. Supertrees have the advantage that they allow the combined analysis of many data sets, each with different taxonomic sampling, thereby circumventing the typical limitation of concatenation approaches. Using a supertree approach on a large data set of 2,807 genes, Pisani *et al.* evaluated the relationship among 127 taxa belonging to the Bacteria (97 species), the Archaea (21 species) and the Eukarya (17 species) (TABLE 1) using an original strategy named 'phylogenetic signal stripping' (REF. 29), which provides a way of identifying discrete phylogenetic signals in a complex data set. The supertree inferred from these 2,807 genes showed the Eukarya as a sister group to the phylum Cyanobacteria, which results from the biased effect of the strong phylogenetic signal provided by genes of chloroplast origin in plants. The authors identified and removed the trees supporting this relationship and produced a second supertree, which revealed a sister relationship between the Eukarya and the class Alphaproteobacteria, consistent with a phylogenetic signal expected for genes of mitochondrial origin. They then identified and removed the trees supporting this relationship, as well as those supporting a sister group relationship between the Eukarya and any bacterial group, which potentially represent HGT from different bacterial sources, and produced a third supertree. This supertree placed the Eukarya within the Archaea, specifically with the Thermoplasmatales (an order in the Euryarchaeota), which the authors interpreted as support for a 2D scenario in which the Eukarya are

Table 1 | Summary of seven recent large-scale phylogenomic analyses

Publication	Number of markers used to infer relationships among the three domains	Taxonomic sampling	Number of amino acid positions used	Method	Model supported
Harris <i>et al.</i> ³¹	50	25 bacteria 1 crenarchaeote 7 euryarchaeotes 3 eukaryotes	Variable, depending on the gene	Single-gene analysis Maximum likelihood Maximum parsimony Distance	3D
Ciccarelli <i>et al.</i> ³⁰	31	150 bacteria 4 crenarchaeotes 14 euryarchaeotes 23 eukaryotes	8,090	Concatenation Maximum likelihood	3D
Yutin <i>et al.</i> ²⁸	136	Variable, depending on the gene	Variable, depending on the gene	Single-gene analysis Maximum likelihood	3D
Rivera & Lake ²⁵	Complete genome of <i>Archaeoglobus fulgidus</i>	2 bacteria 1 crenarchaeote 2 euryarchaeotes 2 eukaryotes	Not applicable	Genome content (conditioned reconstruction)	2D (the Eukarya sister of the Crenarchaeota)
Pisani <i>et al.</i> ²⁹	Data not available	97 bacteria 4 crenarchaeotes 17 euryarchaeotes 17 eukaryotes	Variable depending on gene	Supertree	2D (the Eukarya sister of the Thermoplasmatales)
Cox <i>et al.</i> ²³	45	10 bacteria 3 crenarchaeotes 11 euryarchaeotes 16 eukaryotes	5,521	Concatenation Bayesian Maximum likelihood Maximum parsimony	2D (the Eukarya sister of the Crenarchaeota)
Foster <i>et al.</i> ²⁴	41	8 bacteria 8 crenarchaeotes 2 thaumarchaeotes 6 euryarchaeotes 11 eukaryotes	5,222	Concatenation Bayesian Maximum likelihood Maximum parsimony	2D (the Eukarya sister of a group comprising the Crenarchaeota and the Thaumarchaeota)

2D, 'two primary domains' scenario; 3D, 'three primary domains' scenario.

sister to, or evolved directly from, the Thermoplasmatales, in agreement with previously proposed hypotheses^{11–14}.

The value of this work lies in the application of a supertree approach to the analysis of the relationship between the Archaea and the Eukarya, and in the identification of the candidate archaeal parent to the eukaryote domain. However, closer inspection of the experimental procedure suggests that the results should be treated with caution. For example, stripping the trees supporting relationships between the Eukarya and the Cyanobacteria, the Alphaproteobacteria or the Bacteria left only 53 trees (out of the initial 2,807) that could be used to assess the relationship between the Eukarya and the Archaea. From these 53 trees, a supertree showing the emergence of the Eukarya within the Archaea was obtained only when the most divergent data sets were considered, for each of which up to 90% of the data consisted of poorly conserved regions²⁹. Curiously, although the authors question the analysis of Ciccarelli *et al.*³⁰ because it is based on only 31 genes (dubbed 'the tree of one percent' by Dagan and Martin⁴⁵), it seems that their own main result on the relationship between

the Archaea and the Eukarya was obtained from a similarly small data set that was left over after the stripping process and only when including the most divergent data, which are well-known potential sources of artefact (particularly for deep phylogenies). Moreover, as supertrees extract the topological signal contained in single-gene trees, they are subject to the same concerns about the quality of such signal as have been raised for all single-gene analyses, as mentioned above. Indeed, a hint of a potential reconstruction artefact affecting the main results of Pisani *et al.* is the finding that the Thermoplasmatales seem to be a sister group to the Eukarya but do not fall within the Euryarchaeota in this tree. This is clearly a matter for concern, because the Thermoplasmatales are known to be a late-emerging euryarchaeal lineage^{46,47}. Given this clear artefact affecting the archaeal part of the phylogeny derived by Pisani *et al.*, it is legitimate to question how robust the placement of the Eukarya within the Thermoplasmatales and, more generally, within the Archaea really is in their final supertree.

The most recent studies come full circle in that they use state-of-the-art tree-building techniques on gene concatenations,

thereby avoiding some of the pitfalls of earlier studies. Cox *et al.* analysed the complete proteomes of 40 taxa belonging to the Archaea (14 taxa), the Eukarya (16 taxa) and the Bacteria (10 taxa) (TABLE 1) in order to extract universal orthologous proteins²³. They discarded those present in fewer than 36 taxa, those for which alignments contained less than 60 amino acid positions useful for phylogenetic analysis and those for which the trees strongly supported non-monophyly of the Bacteria or of the Eukarya (that is, genes potentially affected by HGT). At the end of this selection process, only 45 reliable universal orthologous proteins were left for analysis; these were concatenated, leading to a final data set of 5,521 amino acid positions. This analysis specifically addressed concerns that heterogeneity in sequence composition can lead to incorrect trees. Cox and colleagues used two new models of sequence evolution for tree reconstruction, one that allows sequence composition to change over time and among lineages and one that allows sequence composition to change among sites in the alignment. Maximum likelihood and Bayesian analyses resulted in a 2D topology in which the Eukarya



Figure 2 | Comparison of the data sets used in five of the phylogenomic analyses. Clusters of orthologous groups (COGs) corresponding to the protein data sets used in the analyses carried out by Yutin *et al.*²⁸, Harris *et al.*³¹, Ciccarelli *et al.*³⁰, Cox *et al.*²³ and Foster *et al.*²⁴. The 17 light pink circles indicate the markers that were considered to contain sufficient phylogenetic signal to resolve the relationship between the Archaea and the Eukarya in the study by Yutin *et al.*²⁸. The five light yellow circles indicate the markers that were discarded by Ciccarelli *et al.*³⁰ because of likely horizontal gene transfers. The dark blue circle corresponds to four paralogous data sets in the Cox *et al.*²³ study that were removed from the follow-up analysis by Foster *et al.*²⁴. The list of genes used in the analysis by Pisani *et al.*²⁹ is not available, and the data used by Rivera and Lake²⁵ was handled differently (see main text), precluding comparison here.

emerged from within the Archaea, specifically, as a sister group to the Crenarchaeota, thus favouring the eocyte hypothesis²³. This result was recently confirmed by the same authors using a larger sampling of archaea²⁴, including the two available complete genomes from the Thaumarchaeota (a recently proposed archaeal phylum formed by mesophilic or thermophilic ammonia-oxidizing lineages^{48–51}), to address a weakness of their previous study (TABLE 1). In this updated analysis, the Eukarya branched as a sister group to a cluster comprising the Crenarchaeota and the Thaumarchaeota²⁴. It would be interesting to discuss whether this result supports the original eocyte

hypothesis, because the Thaumarchaeota and the Crenarchaeota have substantial genomic differences⁴⁹.

The interest of these two analyses clearly resides in the more realistic models used, which aimed to capture sequence evolution more accurately. However, a drawback of sophisticated evolutionary models is the requirement for the accurate estimation of a large number of parameters. For instance, Cox *et al.* chose to remove from the concatenated alignment those sites displaying no changes or only one change²³. One could ask whether this leads to a bias in the estimation of model parameters and, in turn, impacts the resulting trees. Finally, it is important

to note that, although the new evolutionary models used showed a better fit to the data and have been highlighted as the reason for such novel findings⁵², the authors also obtained a 2D topology with more commonly used and less sophisticated homogeneous models^{23,24}. This indicates that their results may be linked more to the data set used (the genes and the taxonomic sampling) than to the new evolutionary models.

A phylogenomic impasse?

These seven studies do not reach a consensus in support of either the 2D or 3D scenarios, with four supporting the 2D scenario and three favouring the 3D scenario (TABLE 1). A striking observation is that, despite the diversity of methods used for the collection of markers and the different taxonomic samplings, the data sets used in most analyses largely overlap (FIG. 2). This is perhaps unsurprising, given that most of these studies required a bacterial out-group, therefore necessitating the use of universal markers^{23,24,28,30,31}. Thus, despite the very large number of complete genomes used and the various means of data collection, the number of markers that can be used to investigate the relationship between the Archaea and the Eukarya is very small and seems to have reached its limit.

Given the similarity of the data sets used (FIG. 2), what might be the reasons underlying the diversity of the reported results? One possibility is the accuracy of the different methodologies. As mentioned above, gene presence or absence matrices such as those used in conditioned reconstruction²⁴ may simply not be suited to inferring relationships among the three domains of life and, thus, to addressing the exact relationship between the Archaea and the Eukarya. Another point of concern relates to single-gene analyses. The study by Yutin *et al.*²⁸ illustrates the fact that single-gene trees do not contain sufficient signal at such phylogenetic depths, and this likewise holds true for the supertree approach taken by Pisani *et al.*²⁹. Gene concatenations^{23,24,30} are generally expected to be more accurate for studying ancient evolutionary relationships^{34,35}, but these have led to opposing results^{23,24,30}. Other than the different approaches used, the different methods of data set assembly are a potential source of incongruence. For example, in the study by Yutin *et al.*²⁸ both taxonomic sampling and phylogenetic analyses were automated, despite the limited number of markers that were kept for final analysis (136 genes). There is much to be gained from the implementation of

Glossary

COG

Family of homologous proteins constructed by comparing predicted proteins from complete genome sequences.

Distance method

Parametric phylogenetic method that aims to find the tree that minimizes the distance among sequences in a model of sequence evolution.

Domain

One of the three main divisions of life: the Archaea, the Bacteria and the Eukarya.

Horizontal gene transfer

The integration of an exogenous gene into the genome of an organism.

LECA

The most recent ancestor of all present-day eukaryotic lineages.

LUCA

The most recent ancestor of all present-day organisms.

Maximum-likelihood method

Parametric phylogenetic method that aims to maximize the likelihood of a tree; that is, the probability of observing the studied alignment according to the tree topology and to a model of sequence evolution.

Maximum-parsimony method

A non-parametric phylogenetic method that aims to find the set of trees which minimizes the number of evolutionary changes.

Monophyletic group

A group of organisms consisting of an ancestor and its descendants.

Orthologous

Derived from a speciation event.

Paralogous

Derived from a gene duplication event.

Single-gene phylogenetic analysis

Reconstruction of a phylogenetic tree based on the comparison of homologous sequences representing a single gene.

Taxonomic sampling

Sampling of homologous sequences chosen for a phylogenetic analysis from all available sequences.

fully automated procedures, but these procedures can easily become an additional source of error. This is exemplified by the fact that some genes that were considered by Yutin *et al.* to contain a useful phylogenetic signal were discarded in the study by Ciccarelli *et al.* because they were found to be affected by HGT^{28,30} (light yellow circles in FIG. 2). In fact, with such a small number of markers, semi-automated strategies with higher quality control may be preferable. However, even in this case, different methodologies may affect the quality of the multiple alignments used and introduce additional sources of incongruence. For instance, the concatenation by Ciccarelli and colleagues was based on 31 proteins representing 8,090 aligned amino acid positions³⁰, which is 2,569 more positions than were used in the concatenation constructed by Cox *et al.*²³ (TABLE 1), although this data set included many more proteins (45 proteins, of which 21 are also in the Ciccarelli *et al.* data set; FIG. 2). Such a difference may, in part, reflect an inappropriate alignment construction by Ciccarelli *et al.*, as mentioned above, and may explain the opposing results of these two analyses.

An additional important concern regards the interpretation of results. For example, Yutin *et al.* interpreted the support of their data for a 3D scenario as evidence that the Eukarya evolved from an as-yet-unidentified (or extinct) deep-branching archaeal lineage²⁸, essentially turning a 3D scenario into a 2D scenario. In order to draw this conclusion, Yutin *et al.* postulate a relationship between eukaryotes and an archaeal lineage that is neither known nor represented in their data set or their trees, which seems to be an ad hoc solution to rescue the 2D scenario. In fact, there is no reason (other than a subjective interpretation) to favour an unidentified or extinct archaeal lineage as the origin of the Eukarya instead of an unidentified or extinct proto-eukaryotic lineage. The emergence of eukaryotes from an extinct archaeal lineage predating the diversification of the present-day archaea is a possibility that cannot be excluded *a priori* but that also cannot be tested by phylogeny. Following the same reasoning, the 3D scenario could also be compatible with the emergence of the Archaea from an extinct eukaryotic lineage, with the emergence of the Archaea and the Eukarya from an extinct bacterial lineage, with the emergence of the Bacteria from an extinct archaeal lineage, and so on.

Finally, the main source of incongruence probably lies in the fact that the markers used carry very little phylogenetic signal

to be able to firmly resolve the relationship between the Archaea and the Eukarya. In fact, it seems that minor variations in the gene data set, the taxonomic sampling or the method of analysis can cause a marked difference in the results obtained. Therefore, a likely and somewhat despairing scenario for the future might be an endless cycle of reanalysis of very similar, limited sets of universal genes without ever reaching a consensus. Is there a way to escape such a phylogenomic impasse? The answer is probably 'yes', but this will require novel strategies to further exploit the phylogenetic signal contained in complete genomes.

An alternative strategy

The use of universal genes to infer the relationship between the Archaea and the Eukarya, although intuitive, has not yet led to a definitive answer. Therefore, other approaches should be explored. We propose a theoretical two-step phylogenomic approach based on non-universal orthologous markers. The first step is the establishment of a robust rooted phylogeny of the Archaea. The current archaeal tree based on small-subunit rRNA is not very well supported statistically, especially at most basal nodes, leaving a large degree of uncertainty about the evolutionary relationships among lineages^{49,53}. However, several recent phylogenomic analyses have provided a more robust archaeal phylogeny based on large concatenations of protein markers such as components of the translation and transcription apparatus, which are slow evolving and less prone to HGT so are good candidate markers for tracing evolutionary history^{54,55}. These studies have led to an almost fully resolved tree that indicates a clear division between the two major archaeal phyla — the Crenarchaeota and the Euryarchaeota — but also suggests the existence of other phyla: the hyperthermophilic Korarchaeota^{56,57} and the globally distributed mesophilic and thermophilic ammonia-oxidizing Thaumarchaeota^{48–51}. However, most of these studies have used eukaryotic sequences as an out-group, either because the analyses use informational genes, many of which have no orthologues in the Bacteria, or because eukaryotic sequences are more similar to archaeal ones, thereby limiting the possibility of tree reconstruction artefacts. However, if the Eukarya arose from within the Archaea, then eukaryotic sequences cannot be used as a valid out-group. Therefore, in the first step of this alternative phylogenomic approach, a robust rooted archaeal tree must be built by using all orthologous

genes shared between the Bacteria and the Archaea, because the Bacteria are an undisputed out-group to the Archaea (FIG. 3A).

The second step is to carry out a phylogenomic analysis of the set of all orthologous genes shared between the Archaea and the Eukarya, using the tree derived in the first step. This allows 3D and 2D scenarios to be tested in the absence of a bacterial out-group. In fact, if the Eukarya represent a sister lineage to the Archaea (the 3D scenario), they will be located on the same branch as the Bacteria (FIG. 3B). Conversely,

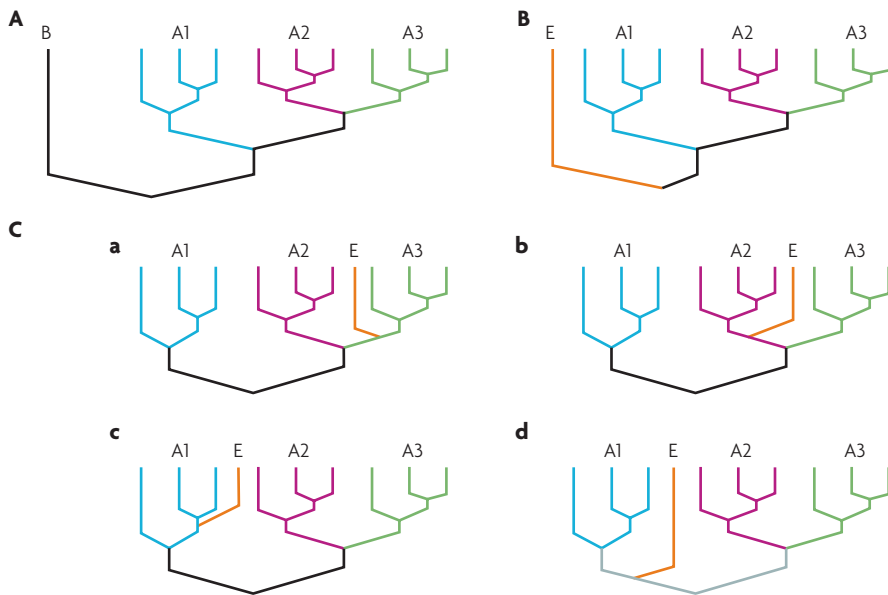


Figure 3 | A two-step phylogenomic approach to investigating the relationship between the Archaea and the Eukarya. **A** | A theoretical reference rooted archaeal phylogeny inferred from orthologous genes that are shared between the Archaea and the Bacteria. For simplicity, only three major archaeal lineages (A1, A2 and A3) are indicated. **B** | The expected placement of eukaryotes with respect to the reference archaeal phylogeny, rooted as in part **A**, in a tree of orthologues shared between the Archaea and the Eukarya, consistent with a 'three primary domains' (3D) hypothesis. **C** | The expected placements of eukaryotes with respect to the reference archaeal phylogeny, rooted as in part **A**, in trees of orthologues shared between the Archaea and the Eukarya, consistent with a 'two primary domains' (2D) hypothesis. Parts **Ca–c** show trees that can discard the 3D hypothesis. Part **Cd** shows the only case for which the approach cannot discriminate between 2D and 3D hypotheses (a tree in which the Eukarya originate from one of the deepest archaeal branches), because the resulting topology is similar to that of the tree in part **B**.

if the Eukarya originated from within the Archaea (the 2D scenario), they will show a branching point different from that of the Bacteria (FIG. 3Ca–c). This strategy will allow the hypothesis that the Eukarya originated from within the Archaea to be definitely excluded, with one exception: if the Eukarya branch as a sister group to either one of the two deepest archaeal branches (FIG. 3Cd, grey branch), the resulting topology will in fact be indistinguishable from that obtained for the 3D scenario (FIG. 3B). Interestingly, this would indicate either that the lineage leading to the present-day eukaryotes emerged very anciently or that the diversification of the Archaea is much more recent than currently assumed, as has been suggested⁵⁸.

This proposed strategy to investigate the evolutionary relationship between the Archaea and the Eukarya has several advantages. First, it increases the size of the data set: the set of proteins shared between the two domains is larger than the set of genes shared universally. Therefore, a stronger phylogenetic signal may be expected than for the very small data set of universal proteins. Using information from only two

domains may also alleviate artefacts that are introduced into universal trees by too high a sequence divergence. Second, because it can exclude the universal gene set, the archaeal–eukaryotic gene set can be used as an independent test of 2D versus 3D scenarios. Third, the observation of eventual incongruence in the internal archaeal phylogeny between the trees obtained from the archaeal–eukaryotic gene set and from the archaeal–bacterial gene set can allow the detection of possible tree reconstruction artefacts. Such an approach will undoubtedly require careful analysis in order to precisely assess the evolutionary history of the markers under study and to extract reliable and unambiguous phylogenetic signal.

Perspectives

A comprehensive understanding of the origin of the Eukarya and their relationship with the Archaea remains an open question and an important challenge in biology. Useful progress in this field has been made, and further progress requires a careful approach: speculation in light of data, known mechanisms and phylogeny should be encouraged

and will allow further insights. For example, support for the 2D eocyte hypothesis needs to be complemented with a well-described biological model, as has been the case for other 2D models^{5,15,59–61} and for the 3D model²⁰. Likewise, phylogenetic testing must follow established guidelines, and care must be taken to identify artefacts and assumptions that may bias the outcome or its interpretation, and to address new difficulties in accurately scaling up analyses.

The Archaea are central to tackling the problem of the origin of the Eukarya, and further exploration of the diversity, biology and evolution of this domain of life will surely yield exciting discoveries in the years to come. For example, new 2D hypotheses could be defined that involve newly proposed major archaeal phyla harbouring specific characteristics, such as the Thaumarchaeota. We encourage biologists working on archaea to interpret their data in an evolutionary framework, as this provides a valuable complement to phylogenomic analyses. Moreover, knowing the exact relationship between the Archaea and the Eukarya is the key to understanding the evolution of the Archaea themselves. In fact, 3D and 2D scenarios put into very different perspectives the emergence of the features shared between these two domains and, therefore, the nature of the archaeal ancestor.

The use of universal genes has not yet provided a definitive answer on the nature of the relationship between the Archaea and the Eukarya — and may perhaps never do so — but we believe that there are grounds to be cautiously optimistic: new ways to exploit the ever-growing amount of genomic data from the three domains of life (such as the analysis of larger data sets of non-universal genes, as proposed here), accompanied by improvements to the evolutionary models and phylogenetic methods to analyse these data, are likely to prove useful for answering this important question.

Simonetta Gribaldo and Patrick Forterre are at the Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Department of Microbiology, 28 rue du Dr Roux, 75015 Paris, France.

Patrick Forterre is also at Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, Bât 409, Université Paris-Sud, Centre d'Orsay, 91405 Orsay Cedex, France.

Anthony M. Poole is at the Department of Molecular Biology and Functional Genomics, Stockholm University, Stockholm SE-106 91, Sweden, and the School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

Vincent Daubin is at the Université de Lyon, Université Lyon 1, Centre National de la Recherche Scientifique, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France.

Céline Brochier-Armanet is at the Laboratoire de Chimie Bactérienne, CNRS UPR9043, 31 chemin Joseph Aiguier, 13402 Marseille, France, and the Université de Provence, 3 place Victor Hugo, 13331 Marseille, France.

e-mails: simonetta.gribaldo@pasteur.fr;
anthony.poole@canterbury.ac.nz;
daubin@biomserv.univ-lyon1.fr; forterre@pasteur.fr;
celine.brochier@ifr88.cnrs-mrs.fr

doi:10.1038/nrmicro2426

- Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579 (1990).
- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl Acad. Sci. USA* **86**, 9355–9359 (1989).
- Gogarten, J. P. *et al.* Evolution of the vacuolar H⁺-ATPase: implications for the origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **86**, 6661–6665 (1989).
- Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623–630 (2006).
- Lopez-Garcia, P. & Moreira, D. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.* **24**, 88–93 (1999).
- Martin, W., Hoffmeister, M., Rotte, C. & Henze, K. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol. Chem.* **382**, 1521–1539 (2001).
- Schleper, C., Jurgens, G. & Jonuscheit, M. Genomic studies of uncultivated archaea. *Nature Rev. Microbiol.* **3**, 479–488 (2005).
- Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
- Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl Acad. Sci. USA* **81**, 3786–3790 (1984).
- Rivera, M. C. & Lake, J. A. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74–76 (1992).
- Margulis, L. *Origin of Eukaryotic Cells* (Yale Univ. Press, New Haven, 1970).
- Margulis, L. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc. Natl Acad. Sci. USA* **93**, 1071–1076 (1996).
- Searcy, D. G., Stein, D. B. & Green, G. R. Phylogenetic affinities between eukaryotic cells and a thermophilic mycoplasma. *Biosystems* **10**, 19–28 (1978).
- Searcy, D. G. Metabolic integration during the evolutionary origin of mitochondria. *Cell Res.* **13**, 229–238 (2003).
- Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
- Moreira, D. & López-García, P. Symbiosis between methanogenic archaea and δ -Proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.* **47**, 517–530 (1998).
- Dagan, T. & Martin, W. Testing hypotheses without considering predictions. *Bioessays* **29**, 500–503 (2007).
- Davidov, Y. & Jurkevitch, E. Comments of Poole and Penny's essay "Evaluating hypotheses for the origin of eukaryotes". *BioEssays* **29**: 74–84. *Bioessays* **29**, 615–616 (2007).
- Poole, A. & Penny, D. Eukaryote evolution: engulfed by speculation. *Nature* **447**, 913 (2007).
- Poole, A. M. & Penny, D. Evaluating hypotheses for the origin of eukaryotes. *Bioessays* **29**, 74–84 (2007).
- Poole, A. M. & Penny, D. Response to Dagan and Martin. *Bioessays* **29**, 611–614 (2007).
- Gribaldo, S. & Philippe, H. Ancient phylogenetic relationships. *Theor. Popul. Biol.* **61**, 391–408 (2002).
- Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaeobacterial origin of eukaryotes. *Proc. Natl Acad. Sci. USA* **105**, 20356–20361 (2008).
- Foster, P. G., Cox, C. J. & Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 2197–2207 (2009).
- Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).
- Lake, J. A. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331**, 184–186 (1988).
- Tourasse, N. J. & Gouy, M. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Mol. Phylogenet. Evol.* **13**, 159–168 (1999).
- Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I. & Koonin, E. V. The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**, 1619–1630 (2008).
- Pisani, D., Cotton, J. A. & McInerney, J. O. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760 (2007).
- Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287 (2006).
- Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. The genetic core of the universal ancestor. *Genome Res.* **13**, 407–412 (2003).
- Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* **462**, 1056–1060 (2009).
- Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
- Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nature Rev. Genet.* **6**, 361–375 (2005).
- Snel, B., Huynen, M. A. & Dutilh, B. E. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* **59**, 191–209 (2005).
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
- Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
- Makarova, K. S., Wolf, Y. I., Sorokin, A. V. & Koonin, E. V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct*, **2**, 35 (2007).
- Snel, B., Bork, P. & Huynen, M. A. Genome phylogeny based on gene content. *Nature Genet.* **21**, 108–110 (1999).
- Fitz-Gibbon, S. T. & House, C. H. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27**, 4218–4222 (1999).
- Lake, J. A. & Rivera, M. C. Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681–690 (2004).
- Deppenmeier, U. *et al.* The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. *J. Mol. Microbiol. Biotechnol.* **4**, 453–461 (2002).
- Ng, W. V. *et al.* Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl Acad. Sci. USA* **97**, 12176–12181 (2000).
- Daubin, V. & Gouy, M. Bacterial molecular phylogeny using supertree approach. *Genome Inform.* **12**, 155–164 (2001).
- Dagan, T. & Martin, W. The tree of one percent. *Genome Biol.* **7**, 118 (2006).
- Gribaldo, S. & Brochier-Armanet, C. The origin and evolution of Archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1007–1022 (2006).
- Boone, D. R. & Castenholz, R. W. *Bergey's Manual of Systematic Bacteriology* (Springer, New York, 2001).
- Fuhrman, J. A., McCallum, K. & Davis, A. A. Novel major archaeobacterial group from marine plankton. *Nature* **356**, 148–149 (1992).
- Brochier-Armanet, C., Boussau, B., Gribaldo, S. & Forterre, P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Rev. Microbiol.* **6**, 245–252 (2008).
- DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl Acad. Sci. USA* **89**, 5685–5689 (1992).
- Hallam, S. J. *et al.* Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl Acad. Sci. USA* **103**, 18296–18301 (2006).
- Zimmer, C. Origins. On the origin of eukaryotes. *Science* **325**, 666–668 (2009).
- Robertson, C. E., Harris, J. K., Spear, J. R. & Pace, N. R. Phylogenetic diversity and ecology of environmental archaea. *Curr. Opin. Microbiol.* **8**, 638–642 (2005).
- Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631–639 (2002).
- Brochier, C., Forterre, P. & Gribaldo, S. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **5**, 36 (2005).
- Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl Acad. Sci. USA* **93**, 9188–9193 (1996).
- Elkins, J. G. *et al.* A korarchaeal genome reveals insights into the evolution of the archaea. *Proc. Natl Acad. Sci. USA* **105**, 8102–8107 (2008).
- Cavalier-Smith, T. The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76 (2002).
- Martin, W. Archaeobacteria (Archaea) and the origin of the eukaryotic nucleus. *Curr. Opin. Microbiol.* **8**, 630–637 (2005).
- Lopez-Garcia, P. & Moreira, D. Selective forces for the origin of the eukaryotic nucleus. *Bioessays* **28**, 525–533 (2006).
- Lopez-Garcia, P. & Moreira, D. On hydrogen transfer and a chimeric origin of eukaryotes. *Trends Biochem. Sci.* **24**, 424 (1999).
- Bapteste, E., Charlebois, R. L., MacLeod, D. & Brochier, C. The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure. *Genome Biol.* **6**, R85 (2005).
- Mans, B. J., Anantharaman, V., Aravind, L. & Koonin, E. V. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* **3**, 1612–1637 (2004).
- Field, M. C. & Dacks, J. B. First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Curr. Opin. Cell Biol.* **21**, 4–13 (2009).
- Embley, T. M. Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 1055–1067 (2006).
- van der Giezen, M. & Tovar, J. Degenerate mitochondria. *EMBO Rep.* **6**, 525–530 (2005).
- Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol. Biol. Evol.* **22**, 1053–1066 (2005).
- Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* **7**, 211–221 (2006).
- Nakamura, T. M. & Cech, T. R. Reversing time: origin of telomerase. *Cell* **92**, 587–590 (1998).
- Ramesh, M. A., Malik, S. B. & Logsdon, J. M. Jr. A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Curr. Biol.* **15**, 185–191 (2005).
- Desmond, E. & NRC-13, S. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol. Evol.* **2009**, 364–381 (2009).
- Eme, L., Moreira, D., Talla, E. & Brochier-Armanet, C. A complex cell division machinery was present in the last common ancestor of eukaryotes. *PLoS ONE* **4**, e5021 (2009).
- Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* **52**, 297–354 (2002).
- Jekely, G. Small GTPases and the evolution of the eukaryotic cell. *Bioessays* **25**, 1129–1138 (2003).
- Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**, 456–502 (1997).
- Lecompte, O., Ripp, R., Thierry, J. C., Moras, D. & Poch, O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* **30**, 5382–5390 (2002).
- Londei, P. Evolution of translational initiation: new insights from the archaea. *FEMS Microbiol. Rev.* **29**, 185–200 (2005).
- Werner, F. Structural evolution of multisubunit RNA polymerases. *Trends Microbiol.* **16**, 247–250 (2008).

79. Bell, S. D. & Jackson, S. P. Mechanism and regulation of transcription in archaea. *Curr. Opin. Microbiol.* **4**, 208–213 (2001).
80. Edgell, D. R. & Doolittle, W. F. Archaea and the origin(s) of DNA replication proteins. *Cell* **89**, 995–998 (1997).
81. Leipe, D. D., Aravind, L. & Koonin, E. V. Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27**, 3389–3401 (1999).
82. Dennis, P. P. & Omer, A. Small non-coding RNAs in Archaea. *Curr. Opin. Microbiol.* **8**, 685–694 (2005).
83. Gaspin, C., Cavaille, J., Erauso, G. & Bachelier, J. P. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *J. Mol. Biol.* **297**, 895–906 (2000).
84. Omer, A. D. *et al.* Homologs of small nucleolar RNAs in archaea. *Science* **288**, 517–522 (2000).
85. Hartung, S. & Hopfner, K. P. Lessons from structural and biochemical studies on the archaeal exosome. *Biochem. Soc. Trans.* **37**, 83–87 (2009).
86. Cubonova, L., Sandman, K., Hallam, S. J., Delong, E. F. & Reeve, J. N. Histones in crenarchaea. *J. Bacteriol.* **187**, 5482–5485 (2005).
87. Reeve, J. N., Sandman, K. & Daniels, C. J. Archaeal histones, nucleosomes, and transcription initiation. *Cell* **89**, 999–1002 (1997).
88. Gribaldo, S. & Cammarano, P. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J. Mol. Evol.* **47**, 508–516 (1998).
89. Samson, R. Y., Obita, T., Freund, S. M., Williams, R. L. & Bell, S. D. A role for the ESCRT system in cell division in archaea. *Science* **322**, 1710–1713 (2008).
90. Lindas, A. C., Karlsson, E. A., Lindgren, M. T., Ettema, T. J. & Bernander, R. A unique cell division machinery in the archaea. *Proc. Natl Acad. Sci. USA* **105**, 18942–18946 (2008).
91. Armengaud, J. *et al.* Identification, purification, and characterization of a eukaryotic-like phosphopantetheine adenylyltransferase (coenzyme A biosynthetic pathway) in the hyperthermophilic archaeon *Pyrococcus abyssi*. *J. Biol. Chem.* **278**, 31078–31087 (2003).
92. Lawson, F. S., Charlebois, R. L. & Dillon, J. A. Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Mol. Biol. Evol.* **13**, 970–977 (1996).
93. Pereto, J., Lopez-Garcia, P. & Moreira, D. Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem. Sci.* **29**, 469–477 (2004).
94. Bapteste, E., Brochier, C. & Boucher, Y. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* **1**, 353–363 (2005).
95. Forterre, P., Gribaldo, S., Gadelle, D. & Serre, M. C. Origin and evolution of DNA topoisomerases. *Biochimie* **89**, 427–446 (2007).
96. Ishino, Y., Komori, K., Cann, I. K. & Koga, Y. A novel DNA polymerase family found in Archaea. *J. Bacteriol.* **180**, 2232–2236 (1998).
97. Desmond, E., Brochier-Armanet, C. & Gribaldo, S. Phylogenomics of the archaeal flagellum: rare horizontal gene transfer in a unique motility structure. *BMC Evol. Biol.* **7**, 106 (2007).
98. Ng, S. Y., Zolghadr, B., Driessen, A. J., Albers, S. V. & Jarrell, K. F. Cell surface structures of archaea. *J. Bacteriol.* **190**, 6039–6047 (2008).
99. Volff, J. N. & Altenbuchner, J. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol. Lett.* **186**, 143–150 (2000).
100. French, S. L., Santangelo, T. J., Beyer, A. L. & Reeve, J. N. Transcription and translation are coupled in Archaea. *Mol. Biol. Evol.* **24**, 893–895 (2007).
101. von Dohlen, C. D., Kohler, S., Alsop, S. T. & McManus, W. R. Mealybug β -proteobacterial endosymbionts contain γ -proteobacterial symbionts. *Nature* **412**, 433–436 (2001).
102. Brinkmann, H. & Philippe, H. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**, 817–825 (1999).
103. Forterre, P. & Philippe, H. Where is the root of the universal tree of life? *Bioessays* **21**, 871–879 (1999).
104. Lopez, P., Forterre, P. & Philippe, H. The root of the tree of life in the light of the covarian model. *J. Mol. Evol.* **49**, 496–508 (1999).
105. Skophammer, R. G., Servin, J. A., Herbold, C. W. & Lake, J. A. Evidence for a Gram-positive, eubacterial root of the tree of life. *Mol. Biol. Evol.* **24**, 1761–1768 (2007).
106. de Crecy-Lagard, V. *et al.* Biosynthesis of wyosine derivatives in tRNA: an ancient and highly diverse pathway in archaea. *Mol. Biol. Evol.* **27**, 2062–2077 (2010).
107. Csuros, M. & Miklos, I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.* **26**, 2087–2095 (2009).

Acknowledgements

The authors thank the Fondation des Treilles for support, along with the four anonymous referees for their valuable comments. A.M.P. is a Royal Swedish Academy of Sciences Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation. V.D. is supported by the grants 'Phylariane', 'Ecogenome' and 'Living Deep' from the French ANR (National Agency for Research). C.B.A. is supported by an Action Thématique et Incitative sur Programme (ATIP) of the French Centre National de la Recherche Scientifique (CNRS).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Simonetta Gribaldo and Patrick Forterre's homepage:
<http://www.pasteur.fr/ip/easysite/go/03b-000033-06b/research/scientific-departments/microbiology/units-and-groups/molecular-biology-of-gene-in-extremophiles>

Anthony M. Poole's homepages:
http://www.molbio.su.se/index.php?option=com_content&view=article&id=716&Itemid=76 | <http://www.biol.canterbury.ac.nz/people/poole.shtml>

Vincent Daubin's homepage:
<http://lbe.univ-lyon1.fr/~Daubin-Vincent.html>

Céline Brochier-Armanet's homepage:
<http://www.frangun.org>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF