The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data

James McInerney1,2, Davide Pisani3 and Mary J. O’Connell4

1Department of Biology, National University of Ireland Maynooth, Co. Kildare, Republic of Ireland
2Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK
3School of Biological Sciences and School of Earth Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
4School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Republic of Ireland

The literature is replete with manuscripts describing the origin of eukaryotic cells. Most of the models for eukaryogenesis are either autogenous (sometimes called slow-drip), or symbiogenic (sometimes called big-bang). In this article, we use large and diverse suites of ‘Omics’ and other data to make the inference that autogeneous hypotheses are a very poor fit to the data and the origin of eukaryotic cells occurred in a single symbiosis.

1. Introduction

The value indeed of an aggregate of characters is very evident in natural history
— Charles Darwin [1, p. 417].

How we think we should depict evolutionary history has, of course, an enormous effect on how we analyse data, how we write analytical software and how we depict the final results. If we feel that the evolutionary history of a dataset has been tree-like, then it is likely that the first, perhaps only, analyses we carry out will be a phylogenetic analysis using software that generates, as an output, a tree. We know of course that human populations do not have a tree-like history; therefore, we are usually disinclined to use software for tree reconstruction to depict these histories [2]. A phylogenetic tree can always be derived based on the complete genomes of two parents and their children. However, we know that the tree will be meaningless, because a tree-like process did not generate the data. What this means is that the outcome of an evolutionary analysis is always contingent on our a priori opinions for how the data have evolved. In some cases, as in the above example, our knowledge of the process that generated the data is good enough to let us unambiguously avoid the use of a particularly poorly fitting model (i.e. a tree) to describe the data (i.e. the relationships between the genomes of two parents and their progeny). However, in most cases, we lack the knowledge to unambiguously reject a model (or a class of models) based on previous observations. In such cases, a better course of action would be to consider a variety of models and ask which fits the data best (if not adequately). In this article, we try to understand the patterns we observe that speak about eukaryotic evolutionary history and we assess the goodness-of-fit between the data collected so far and a variety of models for eukaryotic origins and evolution. We do not limit our model analysis to simple alternative phylogenetic trees, we also include processes that are not tree-like.

One of the most profound changes in evolutionary biology has taken place in the past 10 years. We have had to adjust our thinking on the best means of depicting and analysing evolutionary relationships. Instead of using only phylogenetic trees of organisms as the organizing principle, we must now think also in terms of flows of genetic information. Flows of genetic information can be vertical from ancestor to descendants (following a tree-like process) or horizontal from one lineage to another (following a network-like process) [3]. Horizontal flows can be facilitated by plasmids [4], phage [5], viruses [6], transposons [7], gene transfer...
agents [8], intercellular nanotubes [9], or simply the hybridization of sexual species, followed by re-integration of the hybrid into one of the ancestral species [10,11]. Contemporary genomes are, and extinct genomes were, complex mixtures of genetic mergers [12,13], with the horizontal gene flows being as normal as vertical gene flows. This presents us with a problem if we are restricted to using tree-like processes that only depict vertical gene flows to model the data, as we would not be able to model and understand the impact of horizontal gene flows in evolution.

Phenotypes, particularly in single-celled organisms, coalitions of evolving entities [14] and viruses can only be explained by integrating both a horizontal and vertical view of evolution. In large-scale genome sequencing of bacterial strains, we see thousands of recombination events [15], and indeed, on occasion, explaining phenotypes requires methods that explicitly need to account for and remove vertical transmission signals [16]. One thousand eubacterial genes have flowed into the stem lineage of haloarchaeal Archaeobacteria and identification of this important evolutionary event required the modelling of horizontal, rather than vertical gene flows [17]. Indeed, the impossibility of identifying horizontal flows (even when massive) when using only tree-like models is evident in the history of haloarchaeal studies. One of the first analyses of complete genomes from a broad selection of prokaryotes placed the halophiles as deep-branching Archaeobacteria [18]. The phylogenetic position of this lineage was determined through the interaction of two signals, both present in the genomes of halophiles, one pulling this lineage towards the methanogenic euryarchaeotes, the other pulling the halophiles towards the Eubacteria. The existence of two incongruent signals ultimately caused the halophiles to cluster at the base of the archaeobacterial tree—a phylogenetic position that was not supported by the individual gene trees. A tree-based analysis could not get the correct answer for the placement of archaeobacterial halophiles, because halophile evolution is driven by gene flow based adaptive processes rather than by a pure cladogenetic process. Indeed, archaeobacterial evolution more broadly can only be fully understood when considered as the result of large-scale horizontal gene flows [19] interacting with vertical ones. This is because the origins of most major groups of Archaeabacteria coincide with massive flows of genes from Eubacteria to Archaeabacteria [19,20]. It is becoming increasingly evident that at least some of the genes present in every known organism or lineage (animals included) underwent part of their evolution in a completely different lineage [3]. In the case of animals, obvious examples include all genes of mitochondrial origins (that evolved in the alpha-proteobacterial lineage). To conclude, while one can always force chimerical data onto a tree, model misspecification will ensue. In the best-case scenario, this will simply limit our understanding of the generative process underpinning the data. However, in the worst-case scenario, model misidentification will lead to misleading conclusions, such as the finding of support for monophyletic Archaeabacteria [21], a topology that is no longer supported by the data or by better models (see discussion in §8).

Genomes contain exquisitely detailed information about evolutionary history; however, because of a deep-rooted tradition in the use of tree-like models to explain biological evolution, it was felt that a clear understanding of cladogenesis (to be derived using a phylogenetic tree for a sufficient number of genomes [22]) would have been enough to clarify the early evolution of life on Earth. This narrative remains omnipresent, despite our current understanding of the pervasive role of horizontal gene transfer (HGT) in evolution [14]. A tree is expected; therefore, a tree is sought out.

2. What is the problem with trees in prokaryotic evolution?

Phylogenetic trees have a very precise meaning [23]. Each node on a tree is either a contemporary taxon or a hypothetical ancestral taxon and the linkages indicate vertical inheritance. An evolutionary diagram is a tree if it does not display any reticulations (loops), otherwise it is not a tree. Trees depict evolution as a continuously diverging process and the ramifications on trees indicate specification or duplication events. A more general diagram is a phylogenetic network. These are found in two varieties—networks that display uncertainty and networks that display recombination [24]. Recombination networks also have a restrictive set of assumptions—the data are required to be homologous along its entirety, and reticulations on phylogenetic network graphs indicate places where recombination events between homologous regions have occurred.

More than 10 years ago, Creevey et al. [25] showed that strongly supported but completely conflicting phylogenies existed throughout the prokaryote world. Since then, other studies have confirmed that the majority of prokaryotic genome data cannot fit onto a single phylogenetic tree [26–29] (also see [20]). Only gene flow (both horizontal and vertical) analysis can adequately account for prokaryotic evolution [13,30].

More recently, because of our growing realization that so many molecular sequences are in fact composite entities, we have been advocating a more realistic model of sequence evolution where we allow the merging of both homologous and non-homologous evolving entities [31–33]. In this view, we can analyse situations where very different evolving entities, sometimes from the same evolutionary level (e.g. symbiosis of two organisms, merger of two genes) and sometimes from different levels (e.g. plasmids merging with cells, domains merging with genes) can merge [14]. This is a ‘goods-thinking’ outlook and can have many benefits [12]. We have recently defined $N$-rooted fusion graphs that can instead be used to depict branching relationships with mergings [31,33].

An over-reliance on tree-thinking and tree-based methods [34] has led some researchers to proclaim that we have reached in phylogenetic impasse [35]. However, this is surely premature, as there are a wealth of characters that can be used to provide insights into eukaryogenesis.

3. Different data, different viewpoints on eukaryotic origins?

Most scientists are familiar with the ideas of Karl Popper and his development of falsifiability [36], id est, that the demarcation line between science and non-science is whether or not an idea can be tested, or falsified. However, prior to Popper, a popular philosophy of science was that of William Whewell, who put forward the idea of consilience [37].
Consilience means the ‘jumping together’ of pieces of evidence from disparate kinds of data. Consilience does not mean compromise; rather it is an appeal to look at reciprocal supports from disparate sets of data. Whewell argued that support for a particular theory would be made stronger if multiple lines of evidence that were orthogonal to one another were seen to agree with the same overall theory. Furthermore, if hypothesis X and hypothesis Y both tended to support an overall theory T, then if some new evidence supporting hypothesis X arises, this new evidence not only increases support for the overall theory, but it also strengthens our support for hypothesis Y, even if we have not collected any new data that speaks to hypothesis Y. Specifically with regard to the origins of the eukaryotic cell, we might look to disparate datasets in an effort to see how they agree with an overall theory (if indeed they do at all).

Consilience was the most popular way of viewing science at the time when Darwin, Wallace and colleagues were working out the theory of evolution. Indeed in The origin of species, Darwin noted, ‘The importance, for classification, of trifling characters mainly depends on their being correlated with several other characters of more or less importance’ [1, p. 417].

Here, we argue that some of the hypotheses for the origins of eukaryotes have already been falsified (sensu Popper). In other cases, we rely on an accumulation of evidence from disparate sources for hypotheses that have not yet been rejected by specific observations in order to evaluate which model for the origin of the eukaryotes is better supported (sensu Whewell). Indeed, some hypotheses of eukaryogenesis can be rejected simply because the prediction of the model does not fit with the empirical data, and in some other cases, while no individual piece of evidence is enough to unequivocally define or support a model, when the evidence is taken together its support for the considered hypothesis will become overwhelming. We argue that the only current model that can adequately account for the origin of eukaryotes is the ring of life model of Lake and Rivera [38].

4. What do we know about the last eukaryotic common ancestor?

Eukaryotes presumably did not arise ex nihilo. However, there are considerable disagreements concerning how we might read their traits. In the following sections, we first review what is known about eukaryote traits and how they relate to prokaryote traits. We then ask how these traits map onto the origins of the eukaryotic cell, well. In other cases, we find that eukaryotes have homologous structures with all prokaryotes—membranes, for instance are homologous. In other cases, we find that eukaryotes have proteins or structures that are uniquely eukaryotic. In other cases, we find homologies between eukaryotes and some, but not all prokaryotes, and also we find that there are partial homologies when comparing prokaryotic and eukaryotic features. The detail of the homologies, the organelles and the structures is important. Fortunately, the current wave of ‘omics’ studies is providing more and more data that can speak to eukaryotic origins [39–41]. There are a number of traits that are either uniquely eukaryotic or are present in all eukaryotes. These are the traits that most likely trace back to the first eukaryotic cells.

The last eukaryote common ancestor (LECA) had a nucleus and associated with this nucleus were nuclear pore complexes, some of whose proteins appear to be homologous and some analogous in modern eukaryotes [42,43]. Additionally, comparative analysis of chromatin proteins from protists has provided insights into the early evolutionary history of histone and DNA modification, nucleosome assembly and chromatin-remodelling systems [44]. Many of the individual domains in these eukaryotic systems can be shown to have had bacterial precursors, but today they are found in distinctive regulatory complexes that are unique to the eukaryotes [44]. In general, eukaryotes have linear chromosomes and possess centromeres [45], though prokaryotes like Borrelia burgdorferi also contain linear chromosomes and plasmids [46]. It is unlikely that chromosome linearity is homologous, as there is no suggestion of a sister-group relationship between Borrelia and eukaryotes. Similarly, there have been suggestions of a relationship between Planctomyces and eukaryotes, based on the superficial similarity between the eukaryotic nucleus and the planctomycete membrane that encloses its chromosome. Any direct (or intermediate) relationship between Planctomyces and eukaryotes has been shown to be spurious or analogous [47]. In conclusion, though there is some evidence that the eukaryotic nucleus (excluding the genetic material in this case) has some homologies with eubacterial characters, it is mostly uniquely eukaryotic, with its own independent evolutionary history.

The nucleolus is uniquely eukaryotic. However, careful analysis of the phyletic distribution of the proteome of the nucleolus shows that the nucleolar protein domains confirm the archaebacterial origin of the core machinery for ribosome maturation and assembly, but also reveals substantial eubacterial and eukaryotic contributions to nucleolus evolution [48]. The nucleolus as a whole has no homologue in prokaryotes, though the nucleolus parts have homologies in prokaryotes. These ‘parts’ or ‘goods’ [12,49] form a unique structure in eukaryotes. It appears therefore, that the nucleolus is in some respects chimerical.

Introns are not unique to eukaryotes, but spliceosomal introns are [50]. While group II (self-splicing) introns are commonly found in eukaryotes, Eubacteria and Archaebacteria, spliceosomal introns are only found in eukaryotes. It is thought that spliceosomal introns arose from group II introns early in eukaryotic evolution [50]. Furthermore, intron positions are seen to be conserved across animals, plants and many protists, indicating that spliceosomal introns arose early in eukaryotic evolution and differential loss of introns has been a typical mode of evolution since then [50]. If the suggested relationship between group II introns and spliceosomal introns is confirmed, then this suggests a specific flow of genetic material between eukaryotes and Eubacteria.

Capped and polyadenylated mRNA is also unique to eukaryotes and while both kinds of cellular processes are distinct, there is a common involvement of RNA polymerase II, which is a large macromolecular complex involving both non-coding RNAs and 12–14 proteins. Eukaryotes have three different polymerases involved in transcription, while Eubacteria have only one [51]. There are clear homologies between the RNA polymerase II components of eukaryotes and RNA polymerase proteins in Eubacteria and Archaebacteria, though there are more subunits in common between archaebacterial RNA polymerase and eukaryotic polymerase II (10 subunits in common) than between the eubacterial version and the others.
(five subunits in common with the others). Eukaryote polymerase II has two subunits (Rpb8 and Rpb9) that are unique to eukaryotes.

In terms of other processes that are unique to eukaryotes, the anaphase promoting complex or cyclosome (APC/C) is a major component of the cell cycle and is not found in prokaryotes. In total, 24 out of 37 known APC/C subunits, adaptors/co-activators and main targets, could be inferred to be present in the LECA. For the most part, these components are well conserved in all extant eukaryotic lineages.

Meiosis and mitosis are both eukaryotic features, but two points are interesting about this. Firstly, it is thought that meiosis might be an early invention in eukaryotic evolution [52]; secondly, is the observation of cellular fusion and recombination occurring in Archaea bacteria [53,54]. Cellular fusion and recombination are a long way from being sexual reproduction; however, it is interesting in the light of early meiosis in eukaryotes that prokaryotes can carry out such functions.

One of the most significant differences between prokaryotes and eukaryotes is to be found in cell size and cell volume. Typically, there is a 1000-fold difference in the sizes of the cells in the two groups, though there can be huge variation [55]. Biochemical calculations have also found that the energy requirement per gene is orders of magnitude higher in eukaryotes than in prokaryotes [55], and this additional requirement in terms of energy means that eukaryotes typically need a cellular powerhouse with large amounts of energy-generating membranes. Given the clear relationship between the mitochondrion and alphaproteobacteria, this would only support a model for eukaryote origins that involves a mitochondrial endosymbiosis.

In addition to eukaryote-specific features, eukaryotes share a number of traits that seem more to resemble Archaea bacteria than Eubacteria. For instance, eukaryotes and Archaea bacteria both have histones [56,57], whereas Eubacteria have a non-homologous system of 'histone-like' proteins [58].

Eukaryotes have an elaborate cytoskeleton based on tubulin and actins [59,60]; however, many prokaryotes are known to have homologues of these molecules. FtsZ, which is involved in septum formation during cell division in Eubacteria is found in many eukaryotes as well as Eubacteria and in many archaeobacterial groups, except Crenarchaeota [61]. By contrast, while the Crenarchaeota do not have an FtsZ equivalent, the crenarchaeotal protein crenactin, is related to eukaryotic actin or 'actin-related proteins' and is absent from Euryarchaeota [59]. Actins are structural proteins, often having such diverse functions as providing cell structural support and, in animals, facilitating muscle contraction, for instance [62]. Most recently, it has been shown that a newly discovered group of Archaea bacteria—the 'Lokiarchaeota'—possess an actin variant that is the most closely related prokaryotic actin to eukaryotic actins. This 'lokiactin' has a limited distribution within Archaea bacteria and is specifically found in Archaea bacteria that themselves—using ribosomal protein phylogenies—are placed as sister taxa to eukaryotes. Additionally, a group of proteins that are typically responsible for cell proliferation, the GTPases of the Ras superfamily, are found in Lokiarchaeota and many of these sequences form sister-group relationships with eukaryotes on phylogenetic trees. In addition, the Lokiarchaeota contain '[...]' a primordial version of the eukaryotic ESCRT (endosomal sorting complex required for transport) vesicle trafficking pathway' [63, p. 177]. The ESCRT pathway, which up to now was considered unique to eukaryotes is involved in the formation of multivesicular bodies, which are a trafficking system to bring cargo to the vacuole in eukaryotes. However, the ESCRT system is also known to have other functions such as eukaryotic cell abscission, viral budding, exosome secretion and autophagy [64]. These are not typical features of prokaryotes, so future work will have to properly elucidate their likely function in Lokiarchaeota. It is worth mentioning that Lokiarchaeota are, at the time of writing, only known from metagenomic sequencing [63]. However, the finding of Lokiarchaeota suggests a gene flow between eukaryotes and Archaea bacteria, at least for these key genes.

Mitochondria are a unique feature of eukaryotes, though not all eukaryotes possess mitochondria [65–69]. For some time, it was thought that many lineages of eukaryotes that lacked mitochondria were primitively amitochondriate [70]. This was because they branched deeply in phylogenetic trees of eukaryotes using poorly performing phylogenetic methods [71]. However, using better phylogenetic methods [72] and the discovery of typical mitochondrial genes in the genomes of these amitochondriate eukaryotes [73,74], and furthermore the demonstration that some possessed an organelle that is homologous to the mitochondrion (the hydrogenosome), the 'Archaeaoza' hypothesis has been dismantled [65,66,75]. The implication of this work is that we now know of no primitively amitochondriate eukaryotes [76]. That is to say, we know of no eukaryote genomes that are primitively devoid of a significant gene flow from Eubacteria through mitochondrial acquisition.

To facilitate cell division, Walker A-type cytoskeletal ATPase protein MinD is found widely among Eubacteria and in some eukaryotes, but is absent in Archaea bacteria. In prokaryotes, this protein acts to prevent Z-ring formation anywhere, except for mid-cell [77].

An analysis of 630 genes in eukaryotes with a sister-group relationship to alpha-proteobacteria on phylogenetic trees revealed that many biochemical pathways are completely, or nearly, contained in this cohort. These include the pathways for beta-oxidation, electron transport chain, fructose/manose metabolism and pathways for the synthesis of lipids, biotin, haem and iron–sulfur clusters [78]. In addition, the eukaryotic glyceraldehyde-3-phosphate dehydrogenase [79] is of bacterial origin, and the machinery for iron–sulfur cluster assembly (iron–sulfur clusters are an important part of eukaryotic metalloproteins) is related to alpha-proteobacterial genes [80,81]. The implication from this analysis is that much of eukaryotic metabolism is eu-bacterial-derived.

Analyses of homologies between eukaryotes and prokaryotes have consistently found a disjoint between the relationships that have been inferred [69,82–85] (see supporting information, figure S3, for Alvarez-Ponce et al. [85])—http://archaeobacterial.pnas.org/content/suppl/2013/04/01/121137-1110.DCSupplemental/sapp.pdf). In general, while there are many gene families that are found across all three groups, quite often we find gene families that are found in only two of the three groups (i.e. eukaryotes and Eubacteria, and eukaryotes and Archaea bacteria). Analysis of 'omics' data has shown that the phyletic affiliation of these genes has a very important effect on the cellular role of their proteins.

In gene deletion experiments, it is possible to find out which proteins are essential for a cell to exist. Several such studies have been conducted for Saccharomyces cerevisiae [86].
This allows a test of the association between evolutionary history and function. When eukaryotic genes with prokaryote homologues are partitioned into one class that are homologous with archaeabacterial genes and another class that are homologous to euabacterial genes, we see that the genes that are homologous to Archaebacteria are more likely to be lethal upon deletion, they are more highly expressed than the euabacterial genes, the protein products are significantly more central in protein interaction networks (PINs) and are usually under more selective constraint (as judged by dN/dS ratios) [82,83,85]. In addition, there is an interesting difference between eukaryotic genes with archaeabacterial homologues and those with euabacterial homologues—the euabacterial homologues are more likely to be duplicated in large eukaryotic genomes and are more likely to be lost in small euabacterial genomes. The genes with archaeabacterial homologues are more stable in terms of duplication or loss [85].

Phylogenetic analyses using sophisticated heterogeneous models have consistently placed eukaryote informational genes within archaeabacterial clades [87–89]. Phylogenetic supertrees recognize three main kinds of tree topology—eukaryotes branch with Cyanobacteria, eukaryotes branch with alpha-proteobacteria and eukaryotes branch within Archaebacteria [84]. This supertree analysis found that the gene trees infer many other kinds of relationships, but all these topologies were at a low frequency that was no more than expected by chance [84].

It is clear from the above that there are many eukaryotic signature proteins (ESPs) [69,83,85], and many eukaryotic-specific features (proteins, organelles and processes) that are either unique, highly elaborated compared with their prokaryotic equivalents, or completely absent (e.g. simultaneous transcription and translation), just as there are many euabacterial-specific features (e.g. peptidoglycan cell walls) and archaeabacterial-specific features (e.g. ether-linked lipids, methanogenesis in some groups). If these are considered in isolation, explaining their existence requires hypothesizing that either (a) these features evolved in an isolated derived eukaryotic lineage that engaged in sustained gene genesis, or that (b) these eukaryotic-specific features are ancestral and prokaryotes have lost them as a form of simplifying evolution [90].

### Table 1. Hypothesis as depicted in figure 1 and the corresponding path lengths inferred by these hypotheses. EUB, Eubacteria; EUK, Eukaryota; ARC, Archaebacteria; EOC, Eocyte.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Relationship of path lengths to one another</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) three domains</td>
<td>EUK-ARC ≪ EUK-EUB ≈ ARC-EUB</td>
</tr>
<tr>
<td>(b) eukaryotes early</td>
<td>EUB-ARC ≪ EUK-ARC ≈ EUK-ARC</td>
</tr>
<tr>
<td>(c) eocyte</td>
<td>EUK-EOC &lt; EUK-ARC ≪ EUK-EUB ≈ ARC - EUB</td>
</tr>
<tr>
<td>(d) ring of life</td>
<td>EUB-ARC ≈ EUK-EOC &gt; EUK-ARC ≫ EUK-ARC ≈ EUK-EOC</td>
</tr>
</tbody>
</table>

5. Analysis of path lengths on four different hypotheses

Interpretations of evolutionary history usually involve two components: the data and the model. When looking at deep evolutionary history, it is possible to employ many different kinds of data: fossil, cellular, genomic, among others. The model, on the other hand, is a statement of how evolution might have occurred. There are usually two parts to the model: the branching diagram and the process. The overwhelming majority of branching diagrams in the literature are tree diagrams, and in all likelihood for most purposes these are the correct kinds of diagram to use [91]. However, if evolutionary history of the evolving objects under study has contained introgressive events, then a simple diverging tree cannot be used. To use a tree when the data were not generated by a tree-like process is an error in model selection [92]. There are many network-like alternatives to trees, including phylogenetic networks [93], sequence-sharing networks [32] and N-rooted fusion networks [31,33]. Each of these diagrams contains inferences of flows of evolving entities [24]. In this article, we use the relative distances between the major groups of organisms, as inferred on four different evolutionary scenarios, to ask how well the data that have been collected can map onto these distances. The paths we analyse have been defined elsewhere and we are using the available data, together with Popperian and Whewellian philosophy to analyse how well they fit the data.

Using path length analysis (see table 1 and figure 1), we find that there are significant differences in the times that are estimated since the lineages split. Naturally, HGT will confuse these times for some genes and make coalescence seem more recent than it actually is for some genes and more ancient for other genes. However, all things being equal, the different hypotheses suggest radically different path length ratios. When a long separation time exists, we expect that two taxa will have diverged in most of their features. We expect, for instance, that their PINs will have diverged, that genome contents will have changed and biochemistry will be different. For shorter paths, we expect fewer differences. In addition, we expect that the totality of data will map differently onto the different diagrams. Most importantly, all four hypotheses imply different path lengths: none are simply a re-rooting of another hypothesis.

An analysis of table 1 and figure 1 shows some interesting patterns. The three-domains hypothesis (table 1a and figure 1a) implies that the shortest path on the tree is the one between eukaryotes and Archaebacteria—implying that these should be the most similar organisms, on average. The eukaryotes-early hypothesis (table 1b and figure 1b) shows eukaryotes to be equally distant from both kinds of prokaryote and in both cases this is a long path. The ring of life hypothesis (table 1d and figure 1d) shows eukaryotes equally distant to both kinds of prokaryote, though in this case eukaryotes are closer to the prokaryotes than either is to the other. The eocyte hypothesis (table 1c and figure 1c) suggests a short path separating eukaryotes and a subset of the Archaebacteria—the eocytes, while a longer path
simplificationist)

The eukaryotes-early hypothesis has been seen in several incarnations. Its first incarnation was known as the ‘introns-early’ hypothesis, which speculated that genes in pieces were ancestral (perhaps a left-over from the RNA world that preceded it) and in some lineages, these genes were streamlined in order to become prokaryote [94,95]. Rooted phylogenetic trees also were put forward to support this idea [96,97] and even genome content phylogenies analysed using parsimony [98], the authors having called the last common ancestor of all life ‘incongruously complex’. In this scenario, prokaryotes are seen as ‘superficially and secondarily’ simple [96], with the eukaryotic genomes being a ‘[...] unique cell type that cannot be deconstructed into features inherited directly from Archaea and Bacteria’ [99]. How this irreducible complexity emerged, however, is unclear, unless one assumes a previous unsampled (perhaps extra-terrestrial) history during which this complexity first arose (which would make this hypothesis highly unparsimonious when compared against other much simpler hypotheses), or the existence of a creator. Irrespective of the above, assuming the tree in figure 1b, we see that the longest period of independent evolution is the lineage leading to eukaryotes and we see that Archaebacteria and Eubacteria have a shorter path separating them. Ancestrally, the cell type is eukaryote and so the prokaryote cell structure has evolved from a more complex ancestor and prokaryote evolution is one of predominantly reductional evolution. Certainly, we know that prokaryote genome evolution in particular is biased towards deletions [100], though eukaryotes also manifest this bias.

Following the paths in figure 1b, this would imply that somehow when Archaebacteria and Eubacteria split, the Archaebacteria preferentially took with them or retained the less essential, more lowly expressed genes and genes whose protein products were less central in PINs. Loss of non-essential genes is much more common in PINs. Loss of non-essential genes is not new to evolutionary biology and such a scenario is well known, particularly in symbiont or parasite genomes [101]. However, the topology has an additional implication—that the Eubacteria preferentially took with them or retained the less essential genes, the more lowly expressed genes and the genes whose protein products were less central in PINs, while dispensing with the more essential genes. It is not immediately obvious what selective advantage might accrue from deleting genes that are presumably very important for life. In effect, this

Table: Phylogenetic hypotheses for the highest level relationships of life.

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>inferred path lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Eubacteria, Archaebacteria, eukaryotes</td>
</tr>
<tr>
<td>b</td>
<td>Eubacteria, Eukaryotes-early, Archaebacteria, eukaryotes</td>
</tr>
<tr>
<td>c</td>
<td>Eubacteria, eukaryotes, eocytes, other Archaebacteria</td>
</tr>
<tr>
<td>d</td>
<td>Eubacteria, eukaryotes, eocytes, ring of life</td>
</tr>
</tbody>
</table>

6. Three-domains hypothesis (tree-like, autogenous)

The three-domains model places the eukaryotes as a sister-group to Archaebacteria. This implies a shorter path length, on average between eukaryote and archaebacterial genomes. That is to say, we would expect that eukaryotes and Archaebacteria to have more in common than either has in common with Eubacteria (this assumes a somewhat constant rate of genomic change—in effect clocklike behaviour). Instead, the data reveal that eukaryote genomes tend to have more eubacterial genes than archaebacterial genes [67,69,82,83]. This could still be explained on the Woese model, if those eubacterial homologues were the least likely to be lost throughout evolutionary history. However, this hypothesis is falsified by the observation by Alvarez-Ponce et al. [85] of an association between the likelihood of gene loss (and also of gene duplication) and whether a eukaryote gene is homologous to a eubacterial gene or an archaebacterial gene. In other words, despite the longer path between eukaryotes and Eubacteria in figure 1a, this model suggests that a greater number of dispensable genes were retained than were retained from the more closely related Archaebacteria. This finding makes the three-domains hypothesis highly implausible.

Figure 1. Phylogenetic hypotheses for the highest level relationships of life. On the right are the classes of hypothesis described in the text and on the left are the various inferred path lengths. The colour coding is retained for each hypothesis. (Online version in colour.)
hypothosis expects that Eubacteria were effectively eviscerated, but somehow thrived and expanded. This speculation is the opposite of our expectations using standard evolutionary theories.

In addition to the problem with explaining the eukaryotes-early hypothesis using cellular data, the model fails to account for the absence of fossil eukaryotes before 1.7 billion years ago. The oldest known fossil eukaryote *Shuiyspsphaeridium macroreticulatum* is from the Ruyang group (in China) and is dated at approximately 1.6 to 1.8 Ga [102]. By contrast, the oldest known prokaryotic fossils are twice as old [102]. Assuming that single-celled prokaryotes and eukaryotes have been equally amenable to fossilization over time, there is a question over where the missing fossils have gone.

8. Ecocyte hypothesis (tree-like, partial explanation)

The ecocyte hypothesis was first put forward in 1984 on the basis of the similarities in the ultrastructures of ribosomes found in eukaryotes and some Archaeabacteria [103]. This led Lake and colleagues to propose that there was a sister-group relationship between eukaryotes and a subset of Archaeabacteria, which he called eocytes, and not the relationship that was being suggested from studies of small-subunit ribosomal RNA sequences which placed eukaryotes as the sister-group of all Archaeabacteria. This ecocyte hypothesis spent more than 20 years as the lesser-known alternative to the Woesean tree, but has recently received more attention as better phylogenetic methods have been applied to collections of informational genes [87–89] and supertrees have been built from large collections of orthologues [84]. This hypothesis relates to the specific set of relationships between eukaryotes and Archaeabacteria and centres on whether Archaeabacteria are monophyletic or not. If eukaryotes are ancestral, then it is necessary to explain how sophisticated heterogeneous maximum-likelihood phylogenies place these informational genes within the diversity of Archaeabacteria and reject with significance the hypotheses of monophyletic Archaeabacteria [87,89].

Probably the most conclusive support for a within-Archeabacteria relationship for some genes in eukaryotes has come from the newly discovered ‘Lokiarchaeota’ sequences [63]. These sequences clearly identify a sister-group relationship with eukaryotes for some key genes, including members of the Ras superfamily, eukaryotic actins (lokiactins) and the ESCRT pathway. They provide functional clues to the first eukaryotes and provide a genetic toolbox that could explain the origin of phagotrophy.

9. Ring of life (network, complexificationist)

The ring of life hypothesis places a genomic and cellular merger event at the centre of eukaryogenesis. In this scenario, eukaryotes are not ancient: they are a more recent group than either of the two prokaryotic groups. This scenario also implies that neither prokaryotic group is monophyletic, as the eukaryotes arose from within the group and not as a separate lineage. Furthermore, the expectation from this scenario is that Archaeabacteria (specifically, the eocytes) and the Eubacteria (specifically, the mitochondrial ancestor) made similar contributions to the eukaryote. The implication is that eukaryotes are not simply modified Archaeabacteria or ‘Archaeabacteria with some eubacterial parts’, rather that the origin of eukaryotes corresponds to an egalitarian merger of two distant relatives, and without both parts eukaryotes would not emerge. This is what we mean by ‘similar’: this event would not have occurred unless both were present.

Support for this hypothesis comes from genome content phylogenies [38]. Lake and Rivera [104] have used the method of conditioned reconstruction to produce a collection of phylogenetic trees that are incongruent. However, these trees can be perfectly mapped onto a ring [105]. By taking a ‘signal stripping’ approach, Pisani et al. [84] showed that there were three and only three phylogenetic signals uniting prokaryotes and eukaryotes: one specifically uniting eukaryotes and cyanobacteria, one uniting eukaryotes and alpha-proteobacteria and one uniting eukaryotes and Archaeabacteria. An analysis of chaperone systems in eukaryotes showed that there are a mix of genes with homologies to both Archaeabacteria and Eubacteria and that other phylogenies would place them within the diversity of Archaeabacteria. This is exactly what we see [67,68,87].

Calculations from bioenergetics show that the per-gene energy requirement for eukaryotic genes vastly exceeds the prokaryotic requirement and, therefore, eukaryotes can only exist if they have a ‘powerhouse’ [55]. In addition, eukaryote proteomes show evidence of protein interactions being structured according to phylogenetic affiliation [82,83]: eubacterial proteins preferentially interact with eubacterial chaperones and vice versa was normal in eukaryotes. A simple analysis of the phylogenetic affiliations of eukaryotic genes [69,82,83,85] shows there to be evolutionary affiliations between most functional categories of eukaryotic proteins.

This model would expect and require that some gene phylogenies place eukaryotes within the diversity of Eubacteria and also that other phylogenies would place them within the diversity of Archaeabacteria. This is exactly what we see [67,68,87].

Analyses have not attributed equal roles to archaebacterial and eubacterial homologues of eukaryotes. For instance, archaebacterial homologues are more central in PINs (both closeness centrality and betweenness centrality), more slowly evolving, more likely to be lethal upon deletion, and more likely to be involved in informational processes, among other things [83]. Simply stated, archaebacterial genes seem to be more important in eukaryotic genomes, but eubacterial genes seem to be more flexible and more likely to be duplicated or lost [85]. Given a merger scenario for eukaryogenesis, this would imply that at the start of eukaryogenesis there were two genome equivalents in the merger. Over time, one (the archaebacterial) acted as a more central genome, more involved in informational processes. The eubacterial genome could lose the more important informational genes without a fitness cost if much of the cellular duplication, transcription and translation was being carried out by the archaebacterial proteins. The mitochondrion retained some informational proteins, but the nucleo-cytosolic informational functions were the responsibility of the archaebacterial homologues. By contrast, the less central eubacterial homologues were largely involved in metabolic functions and overall, these genes were more numerous.
10. What is a eukaryote?

Eukaryotes are a secondary lineage of life that originated following a symbiotic event involving a eubacterium and an archaeabacterium. As a consequence, the eukaryotic genome is a complex mixture of archaeabacterial and eubacterial genes that were horizontally shacked into the same cocktail glass. In addition, eukaryotes invented their own genes and cellular structures. The major gene flows into eukaryotes have almost certainly been from an archaeabacterium (the eocyte) that would not be hugely different to modern Archaeabacteria, from a eubacterium that would be placed somewhere within the diversity of modern alpha-proteobacteria, and the third gene flow is to be found in the Archaeplastida and arose from within the diversity of Cyanobacteria. Other prokaryote-to-eukaryote gene flows, though they are known to exist [107] and might have had significant phenotypic effects [108], do not seem to have been formative.

Modern eukaryotes are composed largely of archaeabacterial homologues that are more highly expressed, more central in PINs, more likely to be lethal upon deletion, more slowly evolving and more likely to be informational. By contrast, eukaryotic genes with eubacterial homologues are more likely to be duplicated in larger genomes and more likely to be lost in smaller genomes, they are more likely to be involved in mendelian disease in humans, they are less likely to be lethal upon deletion and are faster evolving than their archaeabacterial counterparts.

The main archaeabacterial gene flow into eukaryotes brought with it the raw materials for the eukaryote cytoskeleton, transcription, vesicle trafficking and some elements of cell division. The eubacterial gene flow brought with it beta-oxidation, electron transport chain, fructose/mannose metabolism and pathways for the synthesis of lipids, biotin, haem and iron–sulfur clusters [78]. The eubacterial gene flow is also likely to have brought group II introns, which evolved into spliceosomal introns [50].

To conclude, it is clear that eukaryotes cannot be correctly defined as ‘derived’ Archaeabacteria, or as ‘derived’ Eubacteria. Indeed, to view eukaryotes as being from either the archaeabacterial or the eubacterial lineages is an over-simplification. Each human is derived equally from both parents. They would not exist without a genetic contribution from both, and it does not matter if they look more like their mother or father, or which surname they carry, if any. The reality is that a human only exists as a consequence of a contribution from both parents. Analogously, eukaryotes are equally eubacterial and archaeabacterial. A taxonomic debate exists in the literature on the early evolution of life, whereby hypotheses have been suggested to be characterizable either as three-domains or two-domains based (2D versus 3D hypotheses). This characterization inherently assumes the existence of a tree-like pattern of evolution, which is misleading. Because eukaryotes arose from both Archaeabacteria and Eubacteria, there are only two (monophyletic) lineages of life: (i) cellular life and (ii) the eukaryotes. Monophyletic eukaryotes are nested within monophyletic life. Eukaryotes make domain-based classifications obsolete and we therefore advocate dismissing the use of this term (which can easily be replaced by the term lineage, for instance) entirely. That is, we advocate a ‘domain-free’ view of the history of life, as debates about whether there should be two domains or three are essentialist and moot.

In a pluralistic view of cellular life on the planet, we can see that the merging of eubacterial genes with archaeabacterial genes gave rise to the halophiles and indeed it made an enormous contribution to the origins of most of the major groups of Archaeabacteria. We see that photosynthesis can only be interpreted as a series of gene flows around the prokaryotic and eukaryotic worlds. We see that eukaryotes have arisen as a consequence of major flows between prokaryotes initially (eukaryogenesis), and later, between a prokaryote group and a eukaryotic group (plastid origins) [84]. Life’s history is complex and we should not try to simplify it to suit our need for orderly nomenclatural systems.

Competing interests. We declare we have no competing interests. Funding. We received no funding for this study.