

REVIEW

The last universal common ancestor between ancient Earth chemistry and the onset of genetics

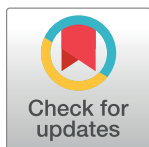
Madeline C. Weiss¹, Martina Preiner¹, Joana C. Xavier¹, Verena Zimorski¹, William F. Martin^{1,2*}

1 Institute of Molecular Evolution, Heinrich Heine University, Düsseldorf, Germany, **2** Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal

* bill@hhu.de

Abstract

All known life forms trace back to a last universal common ancestor (LUCA) that witnessed the onset of Darwinian evolution. One can ask questions about LUCA in various ways, the most common way being to look for traits that are common to all cells, like ribosomes or the genetic code. With the availability of genomes, we can, however, also ask what genes are ancient by virtue of their phylogeny rather than by virtue of being universal. That approach, undertaken recently, leads to a different view of LUCA than we have had in the past, one that fits well with the harsh geochemical setting of early Earth and resembles the biology of prokaryotes that today inhabit the Earth's crust.



OPEN ACCESS

Citation: Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet* 14(8): e1007518. <https://doi.org/10.1371/journal.pgen.1007518>

Editor: Mark Achtman, Warwick Medical School, UNITED KINGDOM

Published: August 16, 2018

Copyright: © 2018 Weiss et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the ERC (666053 <https://erc.europa.eu/>), Volkswagen Foundation (93 046 <https://www.volkswagenstiftung.de/en/funding.html>), and GIF ([1-1321-203.13/2015] <http://www.gif.org.il/Pages/default.aspx>) to WFM. The funders had no role in the preparation of the article.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The very earliest phases of life on Earth witnessed the origin of life and genetics from the elements. There was a time when there was no life on Earth, and there was a time when there were DNA-inheriting cells. The transitions are hard to imagine. Some dates and constraints on the order of events helps us to better grasp the problem. The Earth is 4.5 billion years (Ga) old [1]. By about 4.4 Ga, the moon-forming impact turned the Earth into a ball of boiling lava [1]. Magma oceans with temperatures over 2,000°K forced all water from early accretion into the gas phase and converted all early accreted carbon to atmospheric carbon dioxide (CO₂) [1,2]. By 4.2 to 4.3 Ga, the Earth had cooled sufficiently enough that there was liquid water [3]—those first oceans were about twice as deep as today's [1,2]. Only later, hydrothermal convection currents started sequestering water to the primordial crust and mantle, which today bind one extra ocean volume [4,5]. The first signs of life appear as carbon isotope signatures in rocks 3.95 billion years of age [6]. Thus, somewhere on the ocean-covered early Earth and in a narrow window of time of only about 200 million years, the first cells came into existence. Because the genetic code [7] and amino acid chirality [8] are universal, all modern life forms ultimately trace back to that phase of evolution. That was the time during which the last universal common ancestor (LUCA) of all cells lived.

LUCA, the tree of life, and its roots

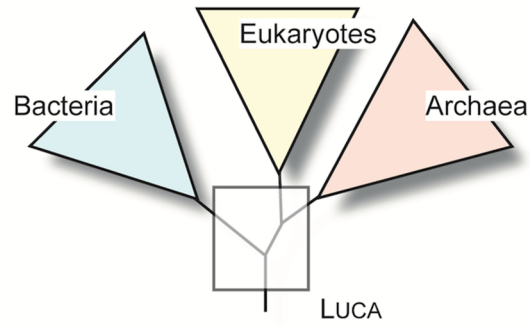
LUCA is a theoretical construct—it might or might not have been something we today would call an organism. It helps to bridge the conceptual gap between rocks and water on the early Earth and ideas about the nature of the first cells. Thoughts about LUCA span decades. Various ideas exist in the literature about how LUCA was physically organized and what properties it possessed. These ideas are traditionally linked to our ideas about the overall tree of life and where its root might lie [9–18]. Phylogenetic trees are, however, ephemeral. It is their inescapable fate to undergo change as new data and new methods of phylogenetic inference emerge. Accordingly, the tree of life has been undergoing a great deal of change of late.

The familiar three-domain tree of life presented by ribosomal RNA [19] depicted LUCA as the last common ancestor of archaea, bacteria, and eukaryotes (Fig 1A). In that framework, efforts to infer the gene content, hence the properties of LUCA, boiled down to identifying genes that were present in eukaryotes, archaea, and bacteria. When the first genomes came out, there were a great many such investigations [20–22], all of which were confronted with the same two recurrent and fundamental problems: 1) How are the three domains related to one another so that gene presence patterns would really trace genes to LUCA as opposed to another evolutionarily more derived branch? 2) Does presence of a gene in two domains (or three) indicate that it was present in the common ancestor of those domains, or could it have reached its current distribution via late invention in one domain and lateral gene transfer (LGT) from one domain to another?

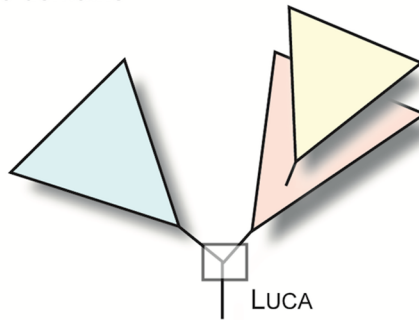
The first problem (the root of the domains) has been the subject of much recent work. Phylogenetic advances and new metagenomic data are changing the three-domain tree [19] into a two-domain tree [24,25]. This is partially a development around phylogenetic methods [24,26–28] but also entails new archaeal lineages that are now being assembled from metagenomic data and that appear to be more closely related to the host that acquired the mitochondrion than any other archaea known so far [29,30]. The two-domain tree showing an "archaeal origin of eukaryotes" [24,28] (Fig 1B) only tells part of the story, though, because eukaryote genomes harbor more bacterial genes than they do archaeal genes by a factor of about 3:1 [31–33], and those bacterial genes furthermore trace to the eukaryote common ancestor [23]. Eukaryotes are not just big, complex archaea; genomically and at the cellular level, they are true chimeras in that they possess archaeal ribosomes in the cytosol and bacterial ribosomes in mitochondria (Fig 1C) [34]. That polarizes cellular evolution in the right direction (there were once debates about eukaryotes being ancestral [10,13,14,22], as discussed elsewhere [35–37]) and identifies eukaryotes as latecomers in evolution, descendants of prokaryotes [38].

Current versions of the two-domain tree focus on the phylogeny of a handful of about 30 genes, mostly for ribosomal proteins (Box 1) but also on sequences from metagenomic samples. The metagenomic studies [29,30] have generated debate. Metagenomic data can bring forth alignments of genes that were sequenced accurately but have the wrong taxonomic label. For example, Da Cunha and colleagues [39] reported that published trees [29] hinge upon a strong signal stemming from one gene out of 30 and that the gene in question (an elongation factor [EF2]) might not be archaeal but eukaryotic instead. Spang and colleagues [40] defended their tree, eliciting more debate [41]. Errors can also occur in the assembly pipeline [42] en route to alignments [43], independent of contamination. Notwithstanding current debate about metagenomics-based trees of life [24,39,40,42,43], we should recall that rRNA itself produces the two-domain tree when various tree construction parameters are employed [24,26,27]. Both data and methods bear upon efforts to construct trees of life. It remains possible that some aspects of domain relationships might never be resolved to everyone's satisfaction—even the endosymbiotic origin of mitochondria is still debated [37]. But the bacterial

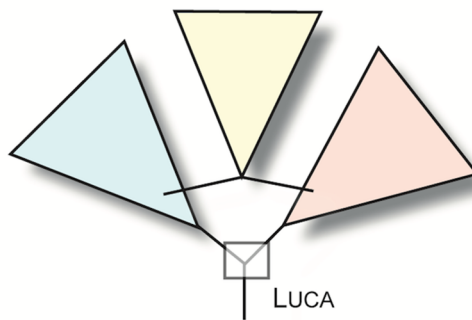
A Three domains



B Two domains



C Symbiosis including mitochondria



D Symbiosis including plastids

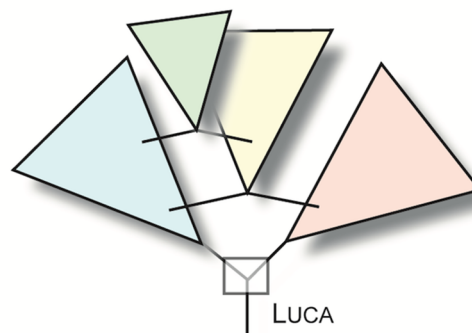


Fig 1. Different views on domain relationships in the tree of life. (A) The three-domain tree: based on rRNA phylogeny, the three domains were of equal rank. (B) The two-domain tree: modern trees show eukaryote cytosolic ribosomes branching within the diversity of archaeal ribosomes. (C) As eukaryotes are not just grownup archaea, the eukaryote ancestor possessed mitochondria. If mitochondrial-derived genes are taken into account, the tree is no longer a bifurcating graph. (D) If plastids are included, the tree becomes even less tree-like because the photosynthetic lineages of eukaryotes also acquired many genes from the plastid ancestor [23].

<https://doi.org/10.1371/journal.pgen.1007518.g001>

origin of mitochondria and their presence in the eukaryote common ancestor [44–47], together with the tendency of eukaryotes to branch within archaeal lineages as archaeal lineage sampling [29,30,48] and phylogenetic methods [24,26,27,32] improve, indicates that eukaryotes arose from prokaryotes and that genes that trace to the common ancestor of archaea and bacteria trace to LUCA.

The second problem (how much LGT has there been between domains) that has impaired progress on LUCA has arguably been more difficult to resolve than the rooting issue. If a given gene is present in bacteria and archaea, was it present in LUCA, or could it have been transferred between domains via LGT? As one important example, early studies pondered the presence of bacterial type oxygen (O₂)-consuming respiratory chains in archaea [21]. Does that

Box 1. The tree of 1% and the tree of everything else

A traditional approach to LUCA has been to simply look for the genes that are present in all genomes. That is easy enough, but the results are sobering. What one finds is a collection of about 30 genes, mostly for ribosomal proteins, telling us that LUCA had a ribosome and had the genetic code, which we already knew [63–65]. That collection of about 30 genes has been in use for about 20 years as concatenated alignments to make trees of lineages based on larger amounts of data than rRNA sequences have to offer [66]. The genes that are present in all lineages (or nearly all) inform us about how LUCA translated mRNA into protein, but they do not tell us about how or where LUCA lived. That information concerns ecophysiology, and physiological traits are not universally conserved—they are what makes microbes different from one another. One can relax the criteria of universal presence a bit and allow for some gene loss in some lineages, in which case, one finds about 100 proteins that are nearly universal [67]. If one puts no size constraints on LUCA's genome and allows loss freely, then all genes present in at least one archaeon and one bacterium trace to LUCA, making it the most versatile organism that ever lived [51]. New insights about microbial phylogeny are emerging from concatenated alignments [24,29,30,42,48,68]. But one has to take care not to get genes from different lineages mixed up, which can be difficult when metagenomes are involved [39,43]. Furthermore, data concatenation has its own pitfalls [66,69,70]. Most modern concatenation studies [29,30,48] employ site-filtering methods in an attempt to remove "noise," but even sites that look "noise free" can still contain bias and conflicting data [63]. Another problem is that popular methods of phylogenetic inference produce inflated confidence intervals on phylogenies and branches [71]. Trees of ca. 30 concatenated proteins are no more immune to phylogenetic error than rRNA is and are prone to additional kinds of error [72]. As it relates to LUCA, regardless of the backbone tree, we still need to know what all proteins say individually about their own phylogenies.

mean that archaea are ancestrally O₂ consumers? As O₂ is the product of cyanobacterial photosynthesis [49] if we presume archaeal O₂ respiration to be an ancestral trait of archaea, it means that archaea arose after cyanobacteria, which are only about 2.5 billion years old and gave rise to plastids (Fig 1D) only about 1.5 billion years ago [50]. If ancestral archaea were oxygen respirers, and ancestral bacteria were too, suddenly neither the two-domain tree nor the three-domain tree (Fig 1) make sense because everything is upside down and rooted in cyanobacteria. Similar issues are encountered for many genes and traits [51]. Lateral gene transfer among prokaryotic domains helps to resolve such problems because it decouples physiology (ecological trait evolution) from phylogeny (ribosomal lineage evolution) [52], but it also makes genes more difficult to trace to LUCA.

Has lateral gene transfer obscured all records?

That takes us to the other extreme. If all genes have been subjected to LGT, as some early claims had it [53], then LUCA would be altogether unknowable from the standpoint of genomes. Early archaeal genomes did indeed uncover abundant transdomain LGT [54], and many bacteria to archaea transfers can be correlated to changes in physiology [55], including the transfer of O₂-consuming respiratory chains [55–58]. For reconstructing LUCA, the issue boils down to determining i) which genes are present in both archaea and bacteria, ii) which of those are present in both prokaryotic domains because of LGT between archaea and bacteria, and iii) which are present because of vertical inheritance from LUCA. For that, there are currently two methodological approaches. One involves making a backbone reference tree from universally conserved genes that are present in each genome—the tree of 1% [59] (see Box 1)—plotting all gene distributions on the tips of that tree, and then estimating which genes trace to LUCA on the basis of various assumed gain and loss parameters [60–62]. If we permit loss freely, many genes will trace back to LUCA; if we assume many gains, LUCA will have few genes [61]. Constraining ancestral genome sizes helps constrain estimates of which genes trace to LUCA [61] but only if we assume that the tree of each gene is compatible with the reference tree, which is a very severe assumption and unlikely to be true. Each gene has its own individual history (Box 1).

Each gene records its own evolutionary history

If any protein-coding genes have been vertically inherited from LUCA, their trees should reflect that. To find such trees, one has to make all trees for all proteins, meaning one has to make clusters for all protein-coding genes from large numbers (thousands) of sequenced genomes. Clusters correspond to "natural" protein families of shared amino acid sequence similarity. Given modern computers, making alignments for all such clusters and making maximum likelihood trees for all such alignments is a tractable undertaking. Because LGT among prokaryotes is a real and pervasive process shaping prokaryote genome evolution [55,58,73–77], one has to treat each gene as a marker of its own evolution, not as a proxy for other genes or as a function that is subordinate to ribosomal phylogeny.

Genes that are present in several bacterial lineages and one archaeal lineage (or vice versa) might have been present in LUCA, but they might also have been the result of LGT [55,56,58]. An example illustrates how each gene tree can discriminate between vertical inheritance from LUCA and interdomain LGT. A recent study investigated the 6.1 million proteins encoded in 1,981 prokaryotic genomes (1,847 bacteria and 134 archaea) [78]. The proteins were clustered using the standard Markov Cluster (MCL) method [79]. The first step in that procedure is a matrix containing 18.5 trillion elements ($(n^2-n)/2$), each element corresponding to a pairwise amino acid sequence comparison. The clustering of such a matrix requires substantial

computational power and is aided by the availability of several terabytes of memory in a single machine. The MCL algorithm samples the distribution of values in the matrix and then starts removing the weak edges, with the value of "weak" being specified by the user. Two kinds of thresholds are typically used in MCL clustering: BLAST e-values and amino acid identity in pairwise alignments.

When the goal of clustering is to make alignments and trees, our group has found that a clustering threshold of 25% amino acid identity is a good rule of thumb. At lower thresholds, amino acid identity starts to approach random values and generates random errors in alignments [80], carrying over as erroneous topologies in trees [81]. That is why Russell F. Doolittle coined the term "twilight zone" for amino acid identity at or below the 20% range [82,83]. Of course, many proteins or domains that clearly share a common ancestry by the measure of related crystal structures do not share more than a random amino acid sequence identity [84]. Such ancient folds will fall into separate clusters at the 25% identity threshold and might thus generate false negatives when it comes to presence in LUCA (but see next section).

From thousands of clusters and trees, a handful remain

Using the 25% identity threshold, the 6.1 million prokaryotic proteins sampled fall into 286,514 clusters of at least two sequences, and 11,093 of those clusters include sequences found in both archaea and bacteria [78]. Many of those clusters involve oxygen-dependent respiratory chains. Did LUCA have 11,000 genes in its genome and breathe oxygen? That is, was LUCA (and hence archaea) descended from cyanobacteria? Neither prospect seems likely enough to warrant further discussion [85]. Knowing that transdomain LGT is prevalent [54–56] and that thousands of typically bacterial genes are shared with only one archaeal group [58], Weiss and colleagues [78] reasoned that a simple way to exclude some LGTs would be to set the minimal phylogenetic criteria that 1) a gene needs to be present in bacteria and archaea, 2) it needs to be present in at least two phylum-level clades, and 3) the tree needs to preserve domain monophyly (Fig 2). Genes that do not fulfil criterion 1 are not candidates for LUCA anyway. The two-phylum-plus-monophyly criteria 2 and 3 make it less likely but not impossible that such a gene attained that distribution via LGT. How so? Criteria 2 and 3 would require one transdomain transfer followed by intradomain transfers to different phyla, while allowing no subsequent, independent transdomain transfers. The last condition is the restrictive one.

Of the 11,093 clusters that harbored sequences in bacteria and archaea, only 355 (3%) passed the simple LGT filter [78]. Put another way, 97% of the sequences present in bacteria and archaea apparently underwent some transdomain LGT, underscoring the degree to which transdomain LGT has influenced gene history since LUCA and underscoring the need to employ phylogenetic filters in search of genes that trace to LUCA [21,51]. The 97% LGT value is important with regard to the 25% clustering threshold and possible false negatives; 97% of all false negatives founded in low-sequence conservation would still not trace to LUCA because of transdomain LGTs. But transdomain LGT has apparently not erased all signals, as 355 genes passed the LGT test, and those genes tell us things about LUCA that we did not know before.

The physiology of LUCA

Most earlier depictions of LUCA focused on what it was like [16]; for example, whether it was like RNA [86], like a virus [87], whether it was like prokaryotes in terms of its genetic code [88], or like eukaryotes in terms of its cellular organization [22]. But traditional approaches lacked information about how and from what LUCA lived [16]. Our phylogenetic approach to LUCA [78] uncovered information about what LUCA was doing: its physiology, its ecology, and its environment. The genes for those physiological traits are not necessarily widespread

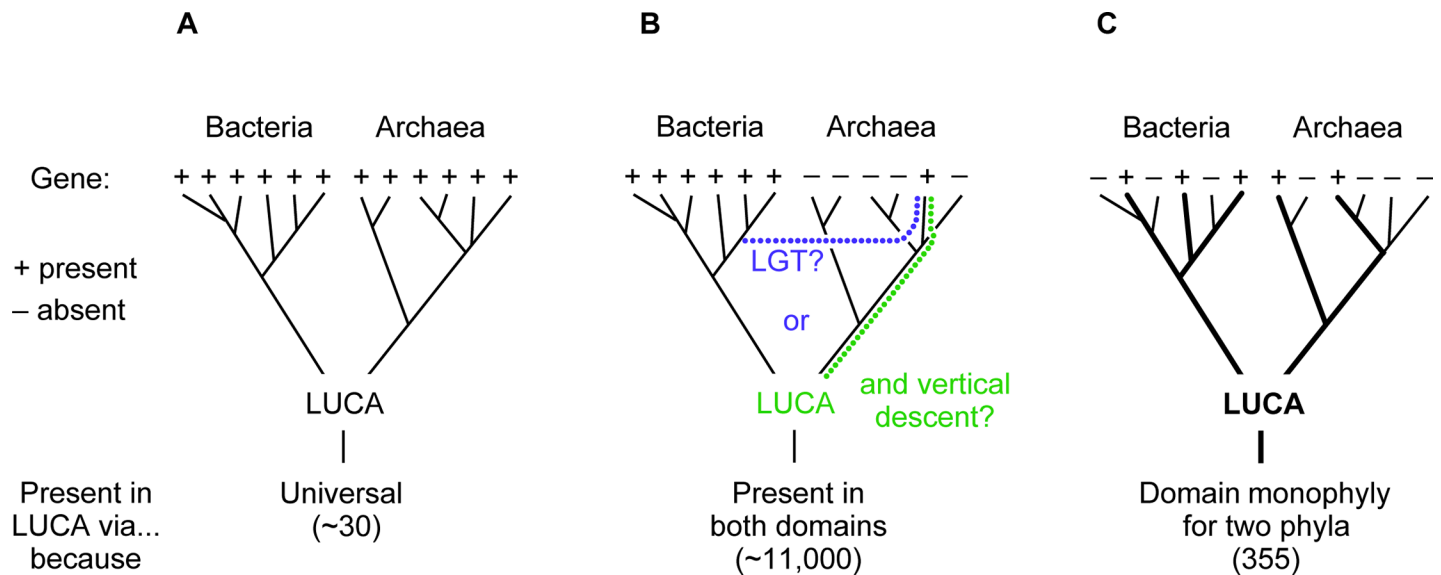


Fig 2. Three ways to infer genes present in LUCA. The gene presence is indicated with a plus sign, absence with a minus sign. a) Genes found universally in both domains, regardless of their tree, trace to LUCA. About 30 fulfil this criterion. b) Another way to trace genes to LUCA is to say that any gene found in both archaea and bacteria was present in LUCA. However, thousands of these genes will have been transferred between bacteria and archaea by LGT so were not necessarily present in LUCA. c) Genes present in only one bacterial or archaeal phylum could easily be the result of LGT and are removed. But presence in two phyla per domain while preserving domain monophyly yields good candidates to have been present in LUCA. Such phylogenies would only result from LGT under very specific and restrictive conditions. They require exactly one transdomain transfer followed by either i) one additional transdomain LGT from the same donor lineage to a different recipient phylum or ii) retention during phylum divergence in the recipient domain, plus—in addition to either criteria i) or ii)—an additional, more subtle but highly restrictive criterion: No further transdomain LGTs occurred during all of evolution. Subsequent transdomain LGT would violate domain monophyly for the gene. Indeed, transdomain LGT is common, and 97% of the trees examined by Weiss and colleagues [78] did not exclude transdomain LGT (remaining 3%, 355 trees, provided in S1 Appendix). LGT, lateral gene transfer; LUCA, last universal common ancestor.

<https://doi.org/10.1371/journal.pgen.1007518.g002>

among modern genomes, but the filtering criteria by Weiss and colleagues [78] only require that these genes are ancient. What Weiss and colleagues [78] found is schematically summarized in Fig 3.

LUCA was an anaerobe, as long predicted by microbiologists [89]. Its metabolism was replete with O₂-sensitive enzymes. These include proteins rich in O₂-sensitive iron-sulfur (FeS) clusters and enzymes that entail the generation of radicals (unpaired electrons) via S-adenosyl methionine (SAM) in their reaction mechanisms. That fits well with the 50-year-old [90] but still modern view that FeS clusters represent very ancient cofactors in metabolism [91–93]. It also fits with newer insights about the ancient and spontaneous (nonenzymatic) chemistry underlying SAM synthesis [94].

LUCA lived from gasses. For carbon assimilation, LUCA used the simplest and most ancient of the six known pathways of CO₂ fixation, called the acetyl-CoA (or Wood-Ljungdahl) pathway [95–97], which is increasingly central for our concepts on early evolution because of its chemical simplicity [97,98] and exergonic nature [99–101]. In the acetyl-CoA pathway, CO₂ is reduced with hydrogen (H₂) to a methyl group and CO. The methyl group is synthesized by the methyl branch of the pathway, which employs different one-carbon (C1) carriers in bacteria (tetrahydrofolate) and archaea (tetrahydromethanopterin), cofactors that are synthesized by unrelated biosynthetic pathways [96]. Carbon monoxide (CO) is synthesized by carbon monoxide dehydrogenase (CODH), the archaeal and bacterial versions of which are distinct but related [96]. The methyl and carbonyl moieties are condensed to an enzyme-bound acetyl group that is removed from a metal cluster in acetyl-CoA synthase (ACS) as an energy rich thioester. Thioesters harbor chemically reactive bonds [102] that play

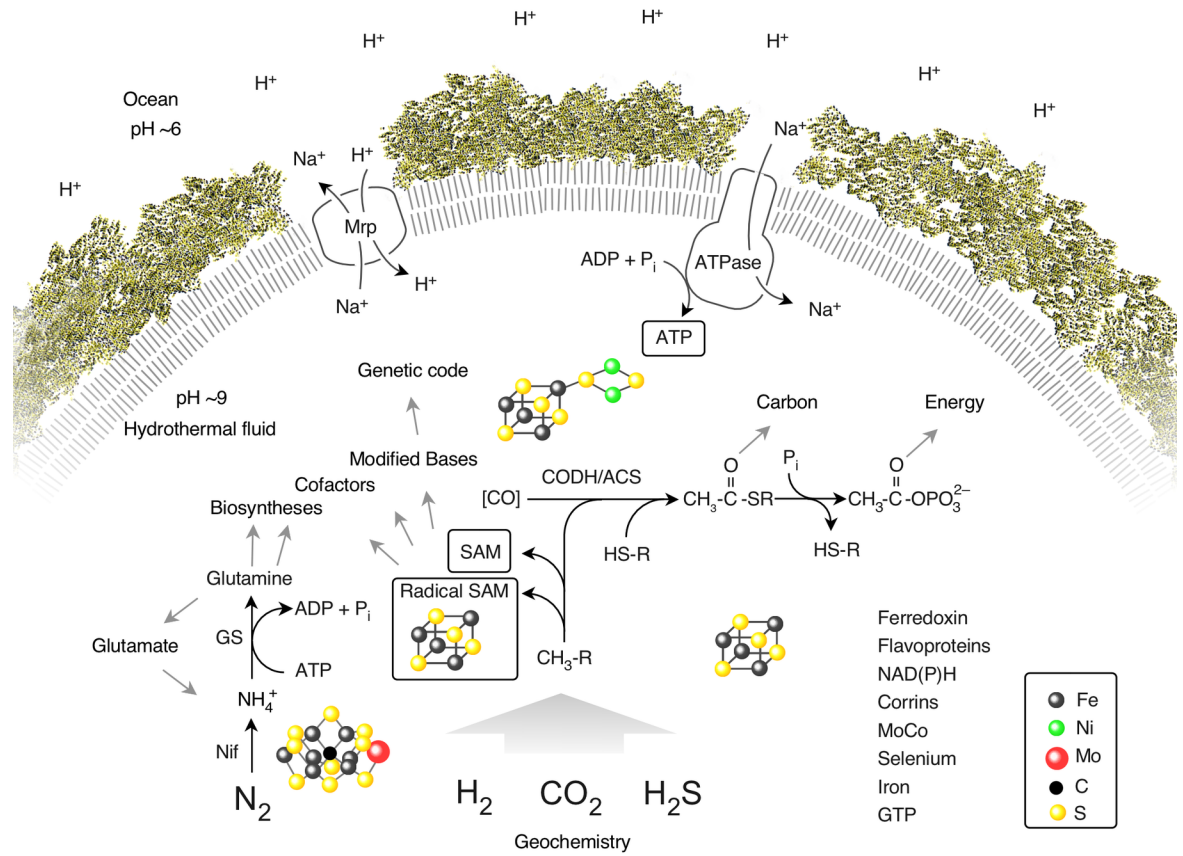


Fig 3. The physiology of LUCA. Summary of the main interactions of LUCA with its environment, reprinted with permission from [78] (supporting trees in S1 Appendix). Components listed at the lower right are present in LUCA. The figure does not make a statement regarding the source of CO in primordial metabolism, symbolized by [CO]. LUCA indisputably possessed genes because it had a genetic code. Transition metal clusters are symbolized. CH₃-R, methyl groups; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase; GS, glutamine synthetase; HS-R, organic thiols; LUCA, last universal common ancestor; Mrp, Mrp type Na⁺/H⁺ antiporter; Nif, nitrogenase; SAM, S-adenosyl methionine.

<https://doi.org/10.1371/journal.pgen.1007518.g003>

a crucial role in energy metabolism [101] and in metabolism in general, both modern and ancient [101,103,104]. Although CODH/ACS clearly does trace to LUCA [78,96], this is not true for the methyl synthesis branch, which consists of unrelated enzymes in bacteria and archaea [78,96].

A recent report [105] argued that the presence of CODH in LUCA did not exclude a heterotrophic lifestyle for LUCA. This argument is problematic because no single enzyme defines a trophic lifestyle. Even Rubisco (D-ribulose-1, 5-bisphosphate carboxylase/oxygenase), the classical Calvin cycle enzyme, is not a marker for autotrophy because Rubisco also functions in a simpler heterotrophic pathway of RNA fermentation [106–108] that is common among archaea and bacteria in marine sediment environments [109]. Moreover, all heterotrophs are derived from autotrophs due to the former requiring the latter as a source of chemically defined growth substrates. The reason is that CO₂ constituted the main carbon source on Earth after the moon-forming impact [1,110], while carbon delivered from space was either too reduced to be fermented (polyaromatic hydrocarbons), too heterogeneous in structure to support microbial growth, or both [108]. Autotrophs with CODH can obtain ATP from CO₂ reduction with H₂ [98,101,110]. Autotrophs without CODH cannot. If we base inferences

about LUCA's lifestyle on broad criteria rather than single genes [105], LUCA was an autotroph [78,108].

Life is about harnessing energy [44]. Thioesters are chemically reactive—they forge direct links between carbon metabolism and energy metabolism (ATP synthesis) as they give rise to acetyl phosphate, the possible precursor of ATP in evolution as a currency of high-energy bonds [111]. Relics of ATP synthesis via acetyl phosphate were found in LUCA's genes [78], as were subunits of the rotor–stator ATP synthase itself. The ATP synthase might appear to present a paradox because no proteins of the proton-pumping machinery that cells use to generate the ion gradient that drives the ATP synthase traced to LUCA [78]. Yet some theories have it that the first cells arose at alkaline hydrothermal vents [91,96,111], meaning that the inside of the vent is more alkaline than the ocean outside. Such naturally existing pH gradients could have been harnessed by LUCA to synthesize ATP (Fig 3). Ancestral ATPases might have harnessed either proton gradients or sodium gradients generated by proton/sodium (H^+/Na^+) dependent antiporters [112], or they might have even been promiscuous for both kinds of ions, similar to the ATPase of modern microbes that live near the thermodynamic limits of life [113].

LUCA's environment was rich in sulfur; thioesters, SAM, proteins rich in FeS and iron–nickel–sulfur (FeNiS) clusters, sulfurtransferases, and thioredoxins were part of its repertoire, as were hydrogenases that could channel electrons from environmental H_2 to reduced ferredoxin, which is the main currency of reducing power (electrons) in anaerobes [114]. A recent report provided phylogenetic evidence that archaea are ancestrally H_2 -dependent methanogens [62], compatible with an autotrophic, H_2 -dependent lifestyle of LUCA.

LUCA had a reverse gyrase, an enzyme typical of thermophiles, suggesting that LUCA liked it hot. But independent of the reverse gyrase, simple chemical kinetics provide strong evidence in favor of a thermophilic origin for the first cells [115,116]. The reason is that only uncatalysed or inorganically catalysed reactions existed before there were enzymes. Their rates of reaction were lower than the enzymatically catalyzed reactions. Between $0^\circ C$ and $120^\circ C$ (the biologically relevant temperature range), organic chemical reaction rates generally increase with temperature [115,116]. Before there were enzymes, high-temperature environments were more conducive to organic chemical reactions than low-temperature environments [115,116]. Taken together, LUCA's requirement for gasses (CO_2 , H_2 , CO, nitrogen [N_2]), the prevalence of sulfide, its affinity to high temperature and metals, plus an ability to use but not generate ion gradients all point to the same environment: alkaline hydrothermal vents.

In addition to shedding light on physiology, the 355 trees that showed domain monophyly (S1 Appendix) [78] also have another interesting property: they are reciprocally rooted. That is, the bacteria are rooted in an archaeal outgroup and vice versa. Genes present in LUCA contain information about their lineages and about the groups of bacteria and archaea that branched most deeply in each domain. In both cases, the answer was clostridia (bacteria) and methanogens (archaea). Those are strictly anaerobic prokaryotes that use the acetyl–CoA pathway; live from CO_2 , H_2 , and CO; fix N_2 ; and today inhabit hydrothermal environments in the Earth's crust [117–119].

The onset of genetics

Though the organization of inanimate matter into living cells with genetics can be charted in mathematical terms [120,121], the biochemical details remain elusive. For example, it is controversial whether LUCA had DNA or not [87]. Several DNA-binding proteins trace to LUCA [78], so it would appear that LUCA possessed DNA, but it is unresolved whether LUCA could

actually replicate DNA. For LUCA, DNA might just have been a chemically stable repository for RNA-based replication [122].

A novel and interesting aspect of LUCA's biology concerns modified bases and the genetic code. Transfer RNA requires modified bases for proper interaction with mRNA (codon–anti-codon wobble base pairing) and with rRNA in the ribosome during translation. That is, modified bases are part of the universal genetic code (Fig 4), which was present in LUCA. Many RNA-modifying enzymes trace to LUCA, particularly the enzymes that modify tRNA. Several of those enzymes are methyltransferases (many SAM dependent), and they remind us that, before the genetic code arose, the four main RNA bases could hardly have been in great supply in pure form because there were no genes or enzymes, only chemical reactions [123]. Spontaneous synthesis of bases in a real early Earth environment like a hydrothermal vent, an environment that lacks the control of a modern laboratory [124], is not likely to generate the four main bases in pure form. Many side products will accumulate, including chemically modified bases [111]. Chemically modified bases from living cells have been reported since the 1970s by pioneering RNA chemists such as Mathias Sprinzl [125] and Henri Grosjean [126]. There are 28 modified bases, mainly occurring in tRNA, that are shared by bacteria and archaea [127]. The modifications are chemically simple, such as the introduction of methyl groups or sulfur and occasionally of acetyl groups and the like (Fig 4).

Chemical modifications in the tRNA anticodon are essential for codon–anticodon interactions to work [128,129]. Modifications of the rRNA are concentrated around the peptidyl transferase site and are also essential for tRNA ribosome interactions [130]. It is possible that the genetic code itself arose in the same chemically reactive environment where LUCA arose and that modified bases in tRNA carry the chemical imprint of that environment [78]. That would forge a link between the early Earth and genetics as we know it. New laboratory syntheses of RNA molecules in the origin of life context now also include investigations of modified bases [131], as it is becoming increasingly clear that these are crucial components at the very earliest phases of molecular and biological evolution.

Moving forward

Investigations of LUCA based on phylogenies of all genes pose new opportunities and new challenges. As environmental sequencing and metagenomics progresses, the number of microbial sequences and new lineages is exploding [48,109]. How will that aspect of metagenomics affect investigations of LUCA? If the criteria for gene age are phylogenetic (prokaryote domain monophyly, presence in at least two bacterial and archaeal “phyla”), then the correct taxonomic assignment of each sequence is very important. A problematic aspect of metagenomic data is that some data handling steps can assign incorrect higher taxon labels to genes [39,41,43], which in turn can falsify phylogenetic relationships. Analyses of cultured microbes or complete genome sequences limit the available sample size but deliver reliable taxon labels, at least at the level of archaea versus bacteria. Clearly, there are trade-offs.

At first sight, LUCA's genome appears doomed to shrinkage. As the sample of complete genomes grows, the list of 355 genes that trace to LUCA by domain monophyly criteria [78] will shrink because each new genome offers new opportunities to uncover recent LGT events for the 355 genes. Recalling that only 3% of the 11,093 clusters investigated [78] appeared free of transdomain LGT, it is evident that the inclusion of new genomes will eventually cause the number 355 to asymptotically approach zero, unless some genes never undergo transdomain LGT, which seems unlikely. What to do? Filtering out recent LGT events would help save LUCA's genome from shrinking to zero. For example, the tree for gene X might violate domain monophyly by one LGT event. If the LGT was recent, affecting members of only one

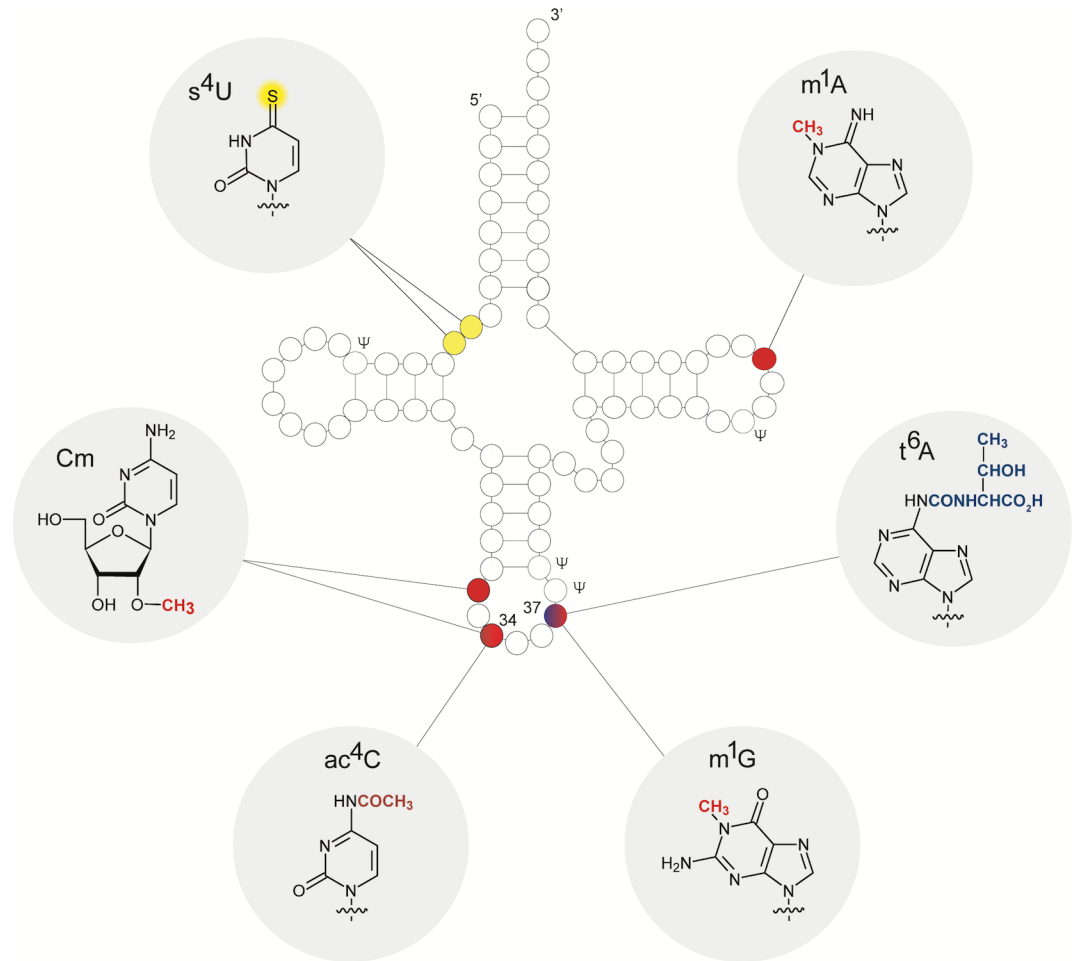


Fig 4. Modified tRNA and nucleoside structures (adapted from [78]). Cloverleaf secondary structure representation of tRNA showing post-transcriptional nucleoside modifications that are conserved among bacteria and archaea in both identity and position. The structures of respective conserved modified nucleosides are highlighted in grey. Methyl and acetyl groups are shown in red and dark red, respectively; sulfur in yellow; and the threonylcarbamoyl group in blue.

<https://doi.org/10.1371/journal.pgen.1007518.g004>

recipient genus or family, it would hardly affect inferences about LUCA, adding gene X to LUCA's list. To identify recent LGTs in prokaryote phylogeny, standard criteria like incomplete amelioration [132], anomalously high-sequence identity [133], or presence in the auxiliary genome [134] will be useful, as will new methods that root unrooted trees [135]. Identifying recent LGTs should allow us to trace more genes to LUCA.

There is also the issue of clustering thresholds to consider, as discussed above. Stringent thresholds produce many small clusters and more relaxed thresholds produce a smaller number of very large clusters [136]. One can argue that large clusters (low stringency) allow one to look further back into time, but they also can generate clusters whose origins trace to duplications in LUCA, in which domain monophyly is violated but not because of LGT. Another factor concerns gene fusions. Genes tend to undergo fusion and fission during evolution [137,138]. In clustering procedures, gene fusions tend to slightly reduce the number of clusters because when they occur, they can bring two fused genes into one alignment, and the weaker phylogenetic signal in the fusion is obscured [23]. Methods to detect fusions exist [139,140]. By detecting gene fusions and dissecting them into their component parts, it might be possible to increase the number of trees that trace to LUCA by phylogenetic criteria.

Investigations into early evolution always elicit protest. For example, there were criticisms [141] of the term "progenote," which Woese and Fox [142] introduced to designate a state of organization below that of a free-living cell [143,144], as shown in Fig 3. In addition, multiple LGTs can, in principle, generate false positives by mimicking vertical inheritance from LUCA [78], but very specific conditions have to be fulfilled (Fig 1C). The challenge is to distill a chronicle of microbial evolution that takes all genes and LGT [145] into account and that conveys information about physiology [146], the energy-releasing reactions that power microbial evolution.

Conclusions

More clues about LUCA's lifestyle are emerging. Investigations of modern biochemical pathways hone in on the same kinds of reactions as the phylogenetic approach [103]. Similarly, laboratory experiments also demonstrate the spontaneous synthesis of end products and intermediates of the acetyl-CoA pathway, the mainstay of LUCA's physiology; new findings show that formate, methanol, acetyl moieties, and even pyruvate arise spontaneously at high yields and at temperatures conducive to life (30°C–100°C) from CO₂, native metals, and water [98,147]. Those conditions are virtually impossible to underbid in terms of chemical simplicity [98], yet they bring forth the core of LUCA's carbon and energy metabolism [78,96,97,101,103] overnight. Did the origin of genetics hinge upon hydrothermal chemical conditions that gave rise to the first biochemical pathways that in turn gave rise to the first cells? Genes that trace to LUCA [78], ancient biochemical pathways [103], and aqueous reactions of CO₂ with iron and water [98,110] all seem to converge on similar sets of simple, exergonic chemical reactions as those that occur spontaneously at hydrothermal vents [148]. From the standpoint of genes, physiology, laboratory chemistry, and geochemistry, it is beginning to look like LUCA was rooted in rocks.

Supporting information

S1 Appendix. ML trees for the 355 protein families that trace to LUCA by phylogenetic criteria. The trees are for the 355 clusters that, after alignment and tree construction, generated ML trees that preserve domain monophyly while also having homologues in ≥ 2 archaeal and ≥ 2 bacterial lineages. These 355 proteins trace to LUCA by those phylogenetic criteria [78]. LUCA, last universal common ancestor; ML, maximum likelihood. (ZIP)

Acknowledgments

We thank Filipa Sousa (Universität Wien), Harun Tüysüz (Max-Planck-Institut für Kohlenforschung, Mülheim an der Ruhr), and Joseph Moran (University of Strasbourg) for discussions.

References

1. Zahnle K, Arndt N, Cockell C, Halliday A, Nisbet E, Selsis F, et al. Emergence of a habitable planet. *Space Sci Rev.* 2007; 129: 35–78.
2. Arndt N, Nisbet E. Processes on the young Earth and the habitats of early life. *Annu Rev Earth Planet Sci.* 2012; 40: 521–549.
3. Mojzsis SJ, Harrison TM, Pidgeon RT. Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4,300 Myr ago. *Nature.* 2001; 409: 178–181. <https://doi.org/10.1038/35051557> PMID: 11196638

4. Hirschmann MM. Water, melting, and the deep earth H₂O cycle. *Annu Rev Earth Planet Sci.* 2006; 34: 629–653.
5. Fei H, Yamazaki D, Sakurai M, Miyajima N, Ohfuji H, Katsura T, et al. A nearly water-saturated mantle transition zone inferred from mineral viscosity. *Sci Adv.* 2017; 3: e1603024. <https://doi.org/10.1126/sciadv.1603024> PMID: 28630912
6. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, et al. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. *Nature.* 2017; 549: 516–518. <https://doi.org/10.1038/nature24019> PMID: 28959955
7. Kubyshev V, Budisa N. Synthetic alienation of microbial organisms by using genetic code engineering: Why and how? *Biotechnol J.* 2017; 12: 1600097–1600110.
8. Haldane JBS. Origin of Life. *The Rationalist Annual.* 1929; 148: 3–10.
9. Crick FHC. The recent excitement in the coding problem. *Prog Nucleic Acid Res Mol Biol.* 1963; 1: 163–217.
10. Doolittle WF, Brown JR. Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci USA.* 1994; 91: 6721–6728. PMID: 8041689
11. Pace NR. Origin of life-facing up to the physical setting. *Cell.* 1991; 65: 531–533. PMID: 1709590
12. Woese C. The universal ancestor. *Proc Natl Acad Sci USA.* 1998; 95: 6854–6859. PMID: 9618502
13. Forterre P, Philippe H. Where is the root of the universal tree of life? *Bioessays.* 1999; 21: 871–879. [https://doi.org/10.1002/\(SICI\)1521-1878\(199910\)21:10<871::AID-BIES10>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1521-1878(199910)21:10<871::AID-BIES10>3.0.CO;2-Q) PMID: 10497338
14. Penny D, Poole A. The nature of the last universal common ancestor. *Curr Opin Genet Dev.* 1999; 9: 672–677. PMID: 10607605
15. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol.* 2003; 1: 127–136. <https://doi.org/10.1038/nrmicro751> PMID: 15035042
16. Becerra A, Delaye L, Islas S, Lazcano A. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annu Rev Ecol Evol Syst.* 2007; 38: 361–379.
17. Di Giulio M. The last universal common ancestor (LUCA) and the ancestors of archaea and bacteria were progenotes. *J Mol Evol.* 2011; 72: 119–126. <https://doi.org/10.1007/s00239-010-9407-2> PMID: 21079939
18. Fox GE. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol.* 2010; 2: a003483. <https://doi.org/10.1101/cshperspect.a003483> PMID: 20534711
19. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA.* 1990; 87: 4576–4579. PMID: 2112744
20. Kyprides N, Overbeek R, Ouzounis C. Universal protein families and the functional content of the Last Universal Common Ancestor. *J Mol Evol.* 1999; 49: 413–423. PMID: 10485999
21. Castresana J, Moreira D. Respiratory chain in the last common ancestor of living organisms. *J Mol Evol.* 1999; 49: 453–460. PMID: 10486003
22. Doolittle WF. The nature of the universal ancestor and the evolution of the proteome. *Curr Opin Struct Biol.* 2000; 10: 355–358. PMID: 10851188
23. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature.* 2015; 524: 427–432. <https://doi.org/10.1038/nature14963> PMID: 26287458
24. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature.* 2013; 504: 231–236. <https://doi.org/10.1038/nature12779> PMID: 24336283
25. McNerney J, Pisani D, O'Connell MJ. The ring of life hypothesis for eukaryote origin is supported by multiple kinds of data. *Philos Trans R Soc Lond B Biol Sci.* 2015; 370: 20140323. <https://doi.org/10.1098/rstb.2014.0323> PMID: 26323755
26. Lake JA. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature.* 1988; 331: 184–186. <https://doi.org/10.1038/331184a0> PMID: 3340165
27. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. *Science.* 1999; 283: 220–221. PMID: 9880254
28. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaeobacterial origin of eukaryotes. *Proc Natl Acad Sci USA.* 2008; 105: 20356–20361. <https://doi.org/10.1073/pnas.0810647105> PMID: 19073919

29. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 2015; 521: 173–179. <https://doi.org/10.1038/nature14447> PMID: 25945739
30. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. 2017; 541: 353–358. <https://doi.org/10.1038/nature21031> PMID: 28077874
31. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*. 2004; 21: 1643–1660. <https://doi.org/10.1093/molbev/msh160> PMID: 15155797
32. Pisani D, Cotton JA, McInerney JO. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*. 2007; 24: 1752–1760. <https://doi.org/10.1093/molbev/msm095> PMID: 17504772
33. Cotton JA, McInerney JO. Eukaryotic genes of archaeobacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc Natl Acad Sci USA*. 2010; 107: 17252–17255. <https://doi.org/10.1073/pnas.1000265107> PMID: 20852068
34. Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. The physiology of phagocytosis in the context of mitochondrial origin. *Microbiol Mol Biol Rev*. 2017; 81: e00008–17. <https://doi.org/10.1128/MMBR.00008-17> PMID: 28615286
35. Forterre P. The common ancestor of Archaea and Eukarya was not an archaeon. *Archaea*. 2013; 372396. <https://doi.org/10.1155/2013/372396> PMID: 24348094
36. Mariscal C, Doolittle WF. Eukaryotes first: How could that be? *Philos Trans R Soc Lond B Biol Sci*. 2015; 370: 20140322. <https://doi.org/10.1098/rstb.2014.0322> PMID: 26323754
37. Harish A, Kurland CG. Mitochondria are not captive bacteria. *J Theor Biol*. 2017; 434: 88–98. <https://doi.org/10.1016/j.jtbi.2017.07.011> PMID: 28754286
38. Dagan T, Roettger M, Bryant D, Martin W. Genome networks root the tree of life between prokaryotic domains. *Genome Biol Evol*. 2010; 2: 379–392. <https://doi.org/10.1093/gbe/evq025> PMID: 20624742
39. Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet*. 2017; 13: e1006810. <https://doi.org/10.1371/journal.pgen.1006810> PMID: 28604769
40. Spang A, Eme L, Saw JH, Caceres EF, Zaremba-Niedzwiedzka K, Lombard J, Guy L, Etema TJG. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet*. 2018; 14: e1007080. <https://doi.org/10.1371/journal.pgen.1007080> PMID: 29596421
41. Da Cunha V, Gaia M, Nasir A, Forterre P. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet*. 2018; 14: e1007215. <https://doi.org/10.1371/journal.pgen.1007215> PMID: 29596428
42. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J–F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 2013; 499: 431–437. <https://doi.org/10.1038/nature12352> PMID: 23851394
43. Williams TA, Embley TM. Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol Evol*. 2014; 6: 474–481. <https://doi.org/10.1093/gbe/evu031> PMID: 24532674
44. Judson OP. The energy expansions of evolution. *Nat Ecol Evol*. 2017; 1: 138. <https://doi.org/10.1038/s41559-017-0138> PMID: 28812646
45. Zachar I, Szathmáry E. Breath-giving cooperation: critical review of origin of mitochondria hypotheses. *Biology Direct*. 2017; 12. <https://doi.org/10.1186/s13062-017-0190-5> PMID: 28806979
46. Lane N. Serial endosymbiosis or singular event at the origin of eukaryotes? *J Theor Biol*. 2017; 434: 58–67. <https://doi.org/10.1016/j.jtbi.2017.04.031> PMID: 28501637
47. Gould SB. Membranes and evolution. *Curr Biol*. 2018; 28: R381–R385. <https://doi.org/10.1016/j.cub.2018.01.086> PMID: 29689219
48. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol*. 2016; 1: 16048. <https://doi.org/10.1038/nmicrobiol.2016.48> PMID: 27572647
49. Fischer WW, Hemp J, Johnson JE. Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci*. 2016; 44: 647–683.
50. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc Natl Acad Sci USA*. 2017; 114: E7737–E7745. <https://doi.org/10.1073/pnas.1620089114> PMID: 28808007
51. Baymann F, Lebrun E, Brugna M, Schoepp-Cothenet B, Giudici-Orticoni MT, Nitschke W. The redox protein construction kit: Pre-last universal common ancestor evolution of energy-conserving enzymes.

- Philos Trans R Soc Lond B Biol Sci. 2003; 358: 267–274. <https://doi.org/10.1098/rstb.2002.1184> PMID: 12594934
52. Martin WF, Bryant DA, Beatty JT. A physiological perspective on the origin and evolution of photosynthesis. *FEMS Microbiol Rev.* 2017; 42: 205–231.
 53. Doolittle WF. Phylogenetic classification and the universal tree. *Science.* 1999; 284: 2124–2129. PMID: [10381871](https://pubmed.ncbi.nlm.nih.gov/10381871/)
 54. Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, et al. The genome of *Methanosarcina mazei*: Evidence for lateral gene transfer between bacteria and archaea. *J Mol Microbiol Biotechnol.* 2002; 4: 453–461. PMID: [12125824](https://pubmed.ncbi.nlm.nih.gov/12125824/)
 55. Wagner A, Whitaker RJ, Krause DJ, Heilers JH, van Wolferen M, van der Does C, et al. Mechanisms of gene flow in archaea. *Nat Rev Microbiol.* 2017; 15: 492–501. <https://doi.org/10.1038/nrmicro.2017.41> PMID: [28502981](https://pubmed.ncbi.nlm.nih.gov/28502981/)
 56. Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, et al. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA.* 2012; 109: 20537–20542. <https://doi.org/10.1073/pnas.1209119109> PMID: [23184964](https://pubmed.ncbi.nlm.nih.gov/23184964/)
 57. López-García P, Zivanovic Y, Deschamps P, Moreira D. Bacterial gene import and mesophilic adaptation in archaea. *Nat Rev Microbiol.* 2015; 13: 447–456. <https://doi.org/10.1038/nrmicro3485> PMID: [26075362](https://pubmed.ncbi.nlm.nih.gov/26075362/)
 58. Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, et al. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature.* 2015; 517: 77–80. <https://doi.org/10.1038/nature13805> PMID: [25317564](https://pubmed.ncbi.nlm.nih.gov/25317564/)
 59. Dagan T, Martin W. The tree of one percent. *Genome Biol.* 2006; 7: 118. <https://doi.org/10.1186/gb-2006-7-10-118> PMID: [17081279](https://pubmed.ncbi.nlm.nih.gov/17081279/)
 60. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. The net of life: Reconstructing the microbial phylogenetic network. *Genome Res.* 2005; 15: 954–959. <https://doi.org/10.1101/gr.3666505> PMID: [15965028](https://pubmed.ncbi.nlm.nih.gov/15965028/)
 61. Dagan T, Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA.* 2007; 104: 870–875. <https://doi.org/10.1073/pnas.0606318104> PMID: [17213324](https://pubmed.ncbi.nlm.nih.gov/17213324/)
 62. Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci USA.* 2017; 114: E4602–E4611. <https://doi.org/10.1073/pnas.1618463114> PMID: [28533395](https://pubmed.ncbi.nlm.nih.gov/28533395/)
 63. Hansmann S, Martin W. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: Influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol.* 2000; 50: 1655–1663. <https://doi.org/10.1099/00207713-50-4-1655> PMID: [10939673](https://pubmed.ncbi.nlm.nih.gov/10939673/)
 64. Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Res.* 2004; 14: 2469–2477. <https://doi.org/10.1101/gr.3024704> PMID: [15574825](https://pubmed.ncbi.nlm.nih.gov/15574825/)
 65. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science.* 2006; 311: 1283–1287. <https://doi.org/10.1126/science.1123061> PMID: [16513982](https://pubmed.ncbi.nlm.nih.gov/16513982/)
 66. Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature.* 1998; 393: 162–165. <https://doi.org/10.1038/30234> PMID: [11560168](https://pubmed.ncbi.nlm.nih.gov/11560168/)
 67. Puigbò P, Wolf YI, Koonin EV. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol.* 2009; 8: 59. <https://doi.org/10.1186/jbiol159> PMID: [19594957](https://pubmed.ncbi.nlm.nih.gov/19594957/)
 68. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature.* 2015; 523: 208–211. <https://doi.org/10.1038/nature14486> PMID: [26083755](https://pubmed.ncbi.nlm.nih.gov/26083755/)
 69. Gadagkar SR, Rosenberg MS, Kumar S. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol.* 2005; 304: 64–74. <https://doi.org/10.1002/jez.b.21026> PMID: [15593277](https://pubmed.ncbi.nlm.nih.gov/15593277/)
 70. Thiergart T, Landan G, Martin WF. Concatenated alignments and the case of the disappearing tree. *BMC Evol Biol.* 2014; 14: 266. <https://doi.org/10.1186/s12862-014-0266-0> PMID: [25547755](https://pubmed.ncbi.nlm.nih.gov/25547755/)
 71. Yang Z, Zhu T. Bayesian selection of misspecified models is over confident and may cause spurious posterior probabilities for phylogenetic trees. *Proc Natl Acad Sci USA.* 2018; 115: 1854–1859. <https://doi.org/10.1073/pnas.1712673115> PMID: [29432193](https://pubmed.ncbi.nlm.nih.gov/29432193/)
 72. Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D. Spectral analysis, systematic bias, and the evolution of chloroplasts. *Mol Biol Evol.* 1999; 16: 573–576.

73. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000; 405: 299–304. <https://doi.org/10.1038/35012500> PMID: 10830951
74. Popa O, Dagan T. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol*. 2011; 14: 615–623. <https://doi.org/10.1016/j.mib.2011.07.027> PMID: 21856213
75. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*. 2011; 7: e1001284. <https://doi.org/10.1371/journal.pgen.1001284> PMID: 21298028
76. McInerney JO, McNally A, O'Connell MJ. Why Prokaryotes have pangenomes. *Nat Microbiol*. 2017; 2: 17040. <https://doi.org/10.1038/nmicrobiol.2017.40> PMID: 28350002
77. Bapteste E, Boucher Y, Leigh J, Doolittle WF. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol*. 2004; 12: 406–411. <https://doi.org/10.1016/j.tim.2004.07.002> PMID: 15337161
78. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, et al. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*. 2016; 1: 16116. <https://doi.org/10.1038/nmicrobiol.2016.116> PMID: 27562259
79. Enright AJ, Van Dongen S, Ouzounis CA. An ancient algorithm for largescale detection of protein families. *Nucleic Acids Res*. 2002; 30: 1575–1584. PMID: 11917018
80. Landan G, Graur D. Heads or tails: A simple reliability check for multiple sequence alignments. *Mol Biol Evol*. 2007; 24: 1380–1383. <https://doi.org/10.1093/molbev/msm060> PMID: 17387100
81. Lockhart PJ, Steel M, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*. 1994; 11: 605. <https://doi.org/10.1093/oxfordjournals.molbev.a040136> PMID: 19391266
82. Doolittle RF. Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. 1st ed. Mill Valley, CA: University Science Books; 1986.
83. Rost B. Twilight zone of protein sequence alignments. *Protein Eng*. 1999; 12: 85–94. PMID: 10195279
84. Rossmann MG, Moras D, Olsen KW. Chemical and biological evolution of nucleotide-binding protein. *Nature*. 1974; 250: 194–199. PMID: 4368490
85. Martin WF, Sousa FL. Early microbial evolution: The age of anaerobes. *Cold Spring Harb Perspect Biol*. 2015; 8: a018127. <https://doi.org/10.1101/cshperspect.a018127> PMID: 26684184
86. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res*. 2002; 30: 1427–1464. PMID: 11917006
87. Forterre P. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA*. 2006; 103: 3669–3674. <https://doi.org/10.1073/pnas.0510333103> PMID: 16505372
88. Di Giulio M. An autotrophic origin for the coded amino acids is concordant with the coevolution theory of the genetic code. *J Mol Evol*. 2016; 83: 93–96. <https://doi.org/10.1007/s00239-016-9760-x> PMID: 27743002
89. Decker K, Jungerman K, Thauer RK. Energy production in anaerobic organisms. *Angew. Chem. Int. Ed*. 1970; 9: 138–158.
90. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*. 1966; 152: 363–366. <https://doi.org/10.1126/science.152.3720.363> PMID: 17775169
91. Russell MJ, Hall AJ. The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J Geol Soc Lond*. 1997; 154: 377–402.
92. Camprubi E, Jordan SF, Vasiliadou R, Lane N. Iron catalysis at the origin of life. *IUBMB Life*. 2017; 69: 373–381. <https://doi.org/10.1002/iub.1632> PMID: 28470848
93. Shock EL, McCollom T, Schulte MD. The emergence of metabolism from within hydrothermal systems. In: Wiegel J, Adams MWW, editors. *Thermophiles: The Keys to Molecular Evolution and the Origin of Life*. Taylor and Francis; 1998. pp. 59–76.
94. Laurino P, Tawfik DS. Spontaneous emergence of S-adenosylmethionine and the evolution of methylation. *Angew Chem Int Ed Engl*. 2017; 56: 343–345. <https://doi.org/10.1002/anie.201609615> PMID: 27901309
95. Ragsdale SW. Enzymology of the Wood-Ljungdahl pathway of acetogenesis. *Ann NY Acad Sci*. 2008; 1125: 129–136. <https://doi.org/10.1196/annals.1419.015> PMID: 18378591
96. Sousa FL, Martin WF. Biochemical fossils of the ancient transition from geoenergetics to bioenergetics in prokaryotic one carbon compound metabolism. *Biochim Biophys Acta*. 2014; 1837: 964–981. <https://doi.org/10.1016/j.bbabi.2014.02.001> PMID: 24513196

97. Fuchs G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu Rev Microbiol.* 2011; 65: 631–658. <https://doi.org/10.1146/annurev-micro-090110-102801> PMID: [21740227](#)
98. Varma SJ, Muchowska KB, Chatelain P, Moran J. Native iron reduces CO₂ to intermediates and end-products of the acetyl-CoA pathway. *Nat Ecol Evol.* 2018; 2: 1019–1024. <https://doi.org/10.1038/s41559-018-0542-2> PMID: [29686234](#)
99. Fuchs G. Variations of the acetyl-CoA pathway in diversely related microorganisms that are not acetogens. In Drake HL, editor. *Acetogenesis*. Chapman & Hall Microbiology Series (Physiology / Ecology / Molecular Biology / Biotechnology). Springer, Boston, MA; 1994. pp. 506–538.
100. Russell MJ, Martin W. The rocky roots of the acetyl-CoA pathway. *Trends Biochem Sci.* 2004; 29: 358–363. <https://doi.org/10.1016/j.tibs.2004.05.007> PMID: [15236743](#)
101. Martin WF, Thauer RK. Energy in ancient metabolism. *Cell.* 2017; 168: 953–955. <https://doi.org/10.1016/j.cell.2017.02.032> PMID: [28283068](#)
102. Semenov SN, Kraft LJ, Ainla A, Zhao M, Baghbanzadeh M, Campbell VE et al. Autocatalytic, bistable, oscillatory networks of biologically relevant organic reactions. *Nature.* 2016; 537: 656–660. <https://doi.org/10.1038/nature19776> PMID: [27680939](#)
103. Goldford JE, Hartman H, Smith TF, Segrè D. Remnants of an ancient metabolism without phosphate. *Cell.* 2017; 168: 1126–1134. <https://doi.org/10.1016/j.cell.2017.02.001> PMID: [28262353](#)
104. Goldford JE, Segrè D. Modern views of ancient metabolic networks. *Curr Opin Syst Biol.* 2018; 8: 117–124. <https://doi.org/10.1016/j.coisb.2018.01.004>
105. Adam PS, Borrel G, Gribaldo S. Evolutionary history of carbon monoxide dehydrogenase/acetyl-CoA synthase, one of the oldest enzymatic complexes. *Proc Natl Acad Sci USA.* 2018; 115: E5836–E5837. <https://doi.org/10.1073/pnas.1716667115> PMID: [29358391](#)
106. Sato T, Atomi H, Imanaka T. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science.* 2007; 315: 1003–1006. <https://doi.org/10.1126/science.1135999> PMID: [17303759](#)
107. Aono R, Sato T, Imanaka T, Atomi H. A pentose bisphosphate pathway for nucleoside degradation in Archaea. *Nat Chem Biol.* 2015; 11: 355–360. <https://doi.org/10.1038/nchembio.1786> PMID: [25822915](#)
108. Schönheit P, Buckel W, Martin WF. On the origin of heterotrophy. *Trends Microbiol.* 2016; 24: 12–24. <https://doi.org/10.1016/j.tim.2015.10.003> PMID: [26578093](#)
109. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell.* 2018; 172: 1181–1197. <https://doi.org/10.1016/j.cell.2018.02.016> PMID: [29522741](#)
110. Sousa FL, Preiner M, Martin WF. Native metals, electron bifurcation, and CO₂ reduction in early biochemical evolution. *Curr Opin Microbiol.* 2018; 43: 77–83. <https://doi.org/10.1016/j.mib.2017.12.010> PMID: [29316496](#)
111. Martin W, Russell MJ. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos Trans R Soc Lond B.* 2007; 362: 1887–1925.
112. Sojo V, Pomiankowski A, Lane N. A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol.* 2014; 12: e1001926. <https://doi.org/10.1371/journal.pbio.1001926> PMID: [25116890](#)
113. Schlegel K, Leone V, Faraldo-Gómez JD, Müller V. Promiscuous archaeal ATP synthase concurrently coupled to Na⁺ and H⁺ translocation. *Proc Natl Acad Sci USA.* 2012; 109: 947–952. <https://doi.org/10.1073/pnas.1115796109> PMID: [22219361](#)
114. Buckel W, Thauer RK. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na⁺ translocating ferredoxin oxidation. *Biochim Biophys Acta.* 2013; 1827: 94–113. <https://doi.org/10.1016/j.bbabi.2012.07.002> PMID: [22800682](#)
115. Wolfenden R, Lewis CA Jr., Yuan Y, Carter CW Jr.. Temperature dependence of amino acid hydrophobicities. *Proc Natl Acad Sci USA.* 2015; 112: 7484–7488. <https://doi.org/10.1073/pnas.1507565112> PMID: [26034278](#)
116. Stockbridge RB, Lewis CA Jr., Yuan Y, Wolfenden R. Impact of temperature on the time required for the establishment of primordial biochemistry, and for the evolution of enzymes. *Proc Natl Acad Sci USA.* 2010; 107: 22102–22105. <https://doi.org/10.1073/pnas.1013647107> PMID: [21123742](#)
117. Chapelle FH, O'Neill K, Bradley PM, Methé BA, Ciufo SA, Knobel LL, et al. A hydrogen-based subsurface microbial community dominated by methanogens. *Nature.* 2002; 415: 312–315. <https://doi.org/10.1038/415312a> PMID: [11797006](#)
118. Lever MA, Heuer VB, Morono Y, Masui N, Schmidt F, Alperin MJ, et al. Acetogenesis in deep subsurface sediments of the Juan de Fuca Ridge Flank: A synthesis of geochemical, thermodynamic, and gene-based evidence. *Geomicrobiol J.* 2010; 27: 183–211.

119. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA*. 1998; 95: 6578–6583. PMID: [9618454](#)
120. Hordijk W, Steel M. Chasing the tail: The emergence of autocatalytic networks. *Biosystems*. 2017; 152: 1–10. <https://doi.org/10.1016/j.biosystems.2016.12.002> PMID: [28027958](#)
121. Steel M, Kauffmann S. A note on random catalytic branching processes. *J Theoret Biol*. 2018; 437: 222–224.
122. Koonin EV, Martin W. On the origin of genomes and cells within inorganic compartments. *Trends Genet*. 2005; 21: 647–654. <https://doi.org/10.1016/j.tig.2005.09.006> PMID: [16223546](#)
123. Baross JA, Martin WF. The ribofilm as a concept for life's origins. *Cell* 2015; 162: 13–15. <https://doi.org/10.1016/j.cell.2015.06.038> PMID: [26140586](#)
124. Sutherland JD. Opinion: Studies on the origin of life—the end of the beginning. *Nat Rev Chem*. 2017; 10: 0012. <https://doi.org/10.1038/s41570-016-0012>
125. Hartmann RK, Mörl M, Sprinzl M. The tRNA world. *RNA*. 2004; 10: 344–349. <https://doi.org/10.1261/ma.5240904> PMID: [14970379](#)
126. Grosjean H, Breton M, Sirand-Pugnet P, Tardy F, Thiaucourt F, Citti C, et al. Predicting the minimal translation apparatus: Lessons from the reductive evolution of mollicutes. *PLoS Genet*. 2014; 10: e1004363. <https://doi.org/10.1371/journal.pgen.1004363> PMID: [24809820](#)
127. Grosjean H, Gupta R, Maxwell ES. Modified nucleotides in archaeal RNAs. In: Blum P, editor. *Archaea: New Models for Prokaryotic Biology*. Norfolk, UK: Caister Academic Press; 2008. pp. 171–196.
128. Yarian C, Townsend H, Czystkowski W, Sochacka E, Malkiewicz AJ, Guenther R, et al. Accurate translation of the genetic code depends on tRNA modified nucleosides. *J Biol Chem*. 2002; 277: 16391–16395. <https://doi.org/10.1074/jbc.M200253200> PMID: [11861649](#)
129. Gustilo EM, Vendeix FA, Agris PF. tRNA's modifications bring order to gene expression. *Curr Opin Microbiol*. 2008; 11: 134–140. <https://doi.org/10.1016/j.mib.2008.02.003> PMID: [18378185](#)
130. Decatur WA, Fournier MJ. rRNA modifications and ribosome function. *Trends Biochem Sci*. 2002; 27: 344–351. PMID: [12114023](#)
131. Schneider C, Becker S, Okamura H, Crisp A, Amatov T, Stadlmeier M, et al. Noncanonical RNA nucleosides as molecular fossils of an early Earth—Generation by prebiotic methylations and carbamoylations. *Angewandte Chemie Int Ed*. 2018; 57: 1–5.
132. Lawrence JG, Ochman H. Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci USA*. 1998; 95: 9413–9417. PMID: [9689094](#)
133. Ku C, Martin WF. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: The 70% rule. *BMC Biol*. 2016; 14: 89. <https://doi.org/10.1186/s12915-016-0315-9> PMID: [27751184](#)
134. Medini D, Donati C, Tettelin H, Massignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev*. 2005; 15: 589–594. <https://doi.org/10.1016/j.gde.2005.09.006> PMID: [16185861](#)
135. Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol*. 2017; 1: 193. <https://doi.org/10.1038/s41559-017-0193> PMID: [29388565](#)
136. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA*. 2008; 105: 10039–10044. <https://doi.org/10.1073/pnas.0800679105> PMID: [18632554](#)
137. Kummerfeld SK, Teichmann SA. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*. 2005; 21: 25–30. <https://doi.org/10.1016/j.tig.2004.11.007> PMID: [15680510](#)
138. Méheust R, Watson AK, Lapointe FJ, Papke RT, Lopez P, Baptiste E. Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. *Genome Biol*. 2018; 19:75. <https://doi.org/10.1186/s13059-018-1454-9>
139. Henry CS, Lerma-Ortiz C, Gerdes SY, Mullen JD, Colasanti R, Zhukov A, et al. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. *BMC Genomics*. 2016; 17: 473. <https://doi.org/10.1186/s12864-016-2782-3> PMID: [27342196](#)
140. Enright AJ, Iliopoulos I, Kyripides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*. 1999; 402: 86–90. <https://doi.org/10.1038/47056> PMID: [10573422](#)
141. Gogarten JP, Deamer D. Is LUCA a thermophilic progenote? *Nat Microbiol*. 2016; 1: 16229. <https://doi.org/10.1038/nmicrobiol.2016.229> PMID: [27886195](#)
142. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA*. 1977; 74: 5088–5090. PMID: [270744](#)

143. Weiss MC, Neukirchen S, Roettger M, Mrnjavac N, Nelson-Sathi S, Martin WF, et al. Reply to 'Is LUCA a thermophilic progenote?' *Nat Microbiol.* 2016; 1: 16230. <https://doi.org/10.1038/nmicrobiol.2016.230> PMID: 27886196
144. Martin WF, Weiss MC, Neukirchen S, Nelson-Sathi S, Sousa FL. Physiology, phylogeny, and LUCA. *Microb Cell.* 2016; 3: 582–587. <https://doi.org/10.15698/mic2016.12.545> PMID: 28357330
145. Martin W. Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *BioEssays.* 1999; 21: 99–104. [https://doi.org/10.1002/\(SICI\)1521-1878\(199902\)21:2<99::AID-BIES3>3.0.CO;2-B](https://doi.org/10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B) PMID: 10193183
146. Liu Y, Beer LL, Whitman WB. Methanogens: A window into ancient sulfur metabolism. *Trends Microbiol.* 2012; 20: 251–258. <https://doi.org/10.1016/j.tim.2012.02.002> PMID: 22406173
147. Muchowska KB, Varma SJ, Chevallot-Beroux E, Lethuillier-Karl L, Li G, Moran J. Metals promote sequences of the reverse Krebs cycle. *Nat Ecol Evol.* 2017; 1: 1716–1721. <https://doi.org/10.1038/s41559-017-0311-7> PMID: 28970480
148. McCollom TM. Abiotic methane formation during experimental serpentinization of olivine. *Proc Natl Acad Sci USA.* 2016; 113: 13965–13970. <https://doi.org/10.1073/pnas.1611843113> PMID: 27821742