nature microbiology

ARTICLES

PUBLISHED: 25 JULY 2016 | ARTICLE NUMBER: 16116 | DOI: 10.1038/NMICROBIOL.2016.116

# The physiology and habitat of the last universal common ancestor

Madeline C. Weiss†, Filipa L. Sousa†, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi and William F. Martin*

**The concept of a last universal common ancestor of all cells (LUCA, or the progenote) is central to the study of early evolution and life's origin, yet information about how and where LUCA lived is lacking. We investigated all clusters and phylogenetic trees for 6.1 million protein coding genes from sequenced prokaryotic genomes in order to reconstruct the microbial ecology of LUCA. Among 286,514 protein clusters, we identified 355 protein families (~0.1%) that trace to LUCA by phylogenetic criteria. Because these proteins are not universally distributed, they can shed light on LUCA's physiology. Their functions, properties and prosthetic groups depict LUCA as anaerobic, $CO_2$-fixing, $H_2$-dependent with a Wood–Ljungdahl pathway, $N_2$-fixing and thermophilic. LUCA's biochemistry was replete with FeS clusters and radical reaction mechanisms. Its cofactors reveal dependence upon transition metals, flavins, *S*-adenosyl methionine, coenzyme A, ferredoxin, molybdopterin, corrins and selenium. Its genetic code required nucleoside modifications and *S*-adenosyl methionine-dependent methylations. The 355 phylogenies identify clostridia and methanogens, whose modern lifestyles resemble that of LUCA, as basal among their respective domains. LUCA inhabited a geochemically active environment rich in $H_2$, $CO_2$ and iron. The data support the theory of an autotrophic origin of life involving the Wood–Ljungdahl pathway in a hydrothermal setting.**

The last universal common ancestor (LUCA) is an inferred evolutionary intermediate[1] that links the abiotic phase of Earth's history with the first traces of microbial life in rocks that are 3.8–3.5 billion years of age[2]. Although LUCA was long considered the common ancestor of bacteria, archaea and eukaryotes[3,4], newer two-domain trees of life have eukaryotes arising from prokaryotes[5,6], making LUCA the common ancestor of bacteria and archaea. Previous genomic investigations of LUCA's gene content have focused on genes that are universally present across genomes[4,7,8], revealing that LUCA had 30–100 proteins for ribosomes and translation. In principle, genes present in one archaeon and one bacterium might trace to LUCA, although their phylogenetic distribution could also be the result of post-LUCA gene origin and interdomain lateral gene transfer (LGT)[8], given that thousands of such gene transfers between prokaryotic domains have been detected[9].

To identify genes that can illuminate the biology of LUCA, we took a phylogenetic approach. Among proteins encoded in sequenced prokaryotic genomes, we sought those that fulfil two simple criteria: (1) the protein should be present in at least two higher taxa of bacteria and archaea, respectively, and (2) its tree should recover bacterial and archaeal monophyly (Fig. 1). Genes meeting both criteria are unlikely to have undergone transdomain LGT, and thus were probably present in LUCA and inherited within domains since the time of LUCA. By focusing on phylogeny rather than universal gene presence, we can identify genes involved in LUCA's physiology—the ways that cells access carbon, energy and nutrients from the environment for growth.

## Results

**Tracing proteins to LUCA by removing transdomain LGTs.** Using the standard Markov cluster algorithm (MCL) at a 25% global identity threshold, we sorted all 6,103,411 protein coding genes in 1,847 bacterial and 134 archaeal genomes into 286,514 protein families, or clusters (see Methods), 11,093 of which contained homologues from bacteria and archaea. After alignment and maximum likelihood (ML) tree construction, only 355 clusters preserve domain monophyly while also having homologues in ≥2 archaeal lineages and ≥2 bacterial lineages (see Methods). Encouragingly, 83% (294/355) of LUCA's genes have some functional annotation (Supplementary Tables 1 and 2), with only a minority belonging to translation.

These 355 proteins were probably present in LUCA and thus provide a glimpse of LUCA's genome. Their distribution across prokaryotic higher taxa is presented in Fig. 2 and Supplementary Fig. 1. Clearly, the list of these 355 genes comes with caveats, such as lineage sampling, sequence conservation and the possibility that multiple LGTs might mimic intradomain vertical inheritance. However, there are also quality benchmarks against which to check the list. For example, LUCA's genes encode 19 proteins involved in ribosome biogenesis and eight aminoacyl tRNA synthetases, which are also essential for the genetic code to work (Supplementary Table 2). Thus, our phylogenetic criteria do not miss the informational core, which itself can be affected by LGTs, such that only subsets of even universally present genes will also meet the domain monophyly criterion. As another benchmark, our phylogenetic criteria return a highly non-random sample of genes. The distribution of functional categories represented among the 355 genes tracing to LUCA is significantly different ($P << 1 \times 10^{-16}$) from that represented in the 11,093 cluster sample (Supplementary Table 3; see Methods), with oxygen sensitive enzymes (Supplementary Table 2) and FeS proteins (Supplementary Table 1) overrepresented in LUCA's list.

**LUCA's microbial ecology reconstructed from genomes.** Reconstructed from genomic data, LUCA emerges as an anaerobic

Institute of Molecular Evolution, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany. †These authors contributed equally to this work. *e-mail: bill@hhu.de



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

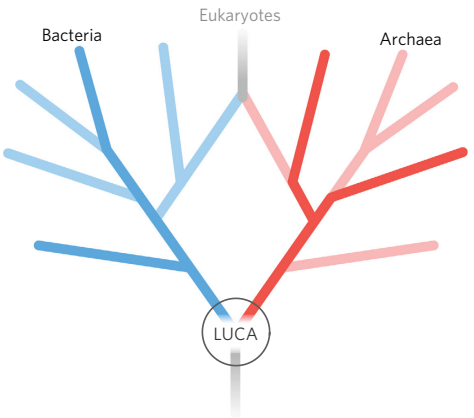


**Figure 1 | Phylogeny for LUCA's genes.** In the two-domain tree of life[5,6], eukaryotes stem from prokaryotes, so the last universal common ancestor, LUCA, is the ancestor of archaea and bacteria. The tree shows a schematic phylogeny of phyla for a gene present in two archaeal and two bacterial phyla and in which both prokaryotic domains are monophyletic. By applying the criteria—(1) the gene should be present in at least two members each of two bacterial phyla and two archaeal phyla (see Methods) and (2) the protein tree should recover monophyly of bacteria and archaea—355 clusters were identified that trace to LUCA.

autotroph[10] that used a Wood–Ljungdahl (WL) pathway[11] and existed in a hydrothermal setting[12,13], but that was only half-alive and was dependent upon geochemistry, as summarized in Fig. 3. LUCA's genes harbour traces of carbon, energy and nitrogen metabolism. Cells conserve energy via chemiosmotic coupling[14] with rotor–stator-type ATP synthases or via substrate-level phosphorylation (SLP)[15]. LUCA's genes encompass components of two enzymes of energy metabolism: phosphotransacetylase (PTA) and an ATP synthase subunit (Supplementary Table 2). PTA generates acetylphosphate from acetyl-CoA, conserving the energy in the thioester bond as the energy-rich anhydride bond of acetylphosphate, which can phosphorylate ADP or other substrates[15]. The PTA reaction plays a central role in autotrophic theories of microbial origins that focus on thioester-dependent SLP as the ancestral state of microbial energy metabolism[16,17]. The presence of a rotor–stator ATP synthase subunit points to LUCA's ability to harness ion gradients for energy metabolism[17], yet the rotor–stator ATP synthase has undergone transdomain LGT[18], excluding many of its subunits from LUCA's set. Crucially, components of electron-transfer-dependent ion-pumping are altogether lacking among LUCA's genes. LUCA's ATPase was possibly able to harness geochemically derived ion gradients[17] via $H^+/Na^+$ antiporters[19], which are present among the membrane proteins in the list (Supplementary Table 2). The presence of reverse gyrase, an enzyme specific for hyperthermophiles[20], indicates a thermophilic lifestyle for LUCA.

Enzymes of chemoorganoheterotrophy are lacking, but enzymes for chemolithoautotrophy are present. Among the six known pathways of $CO_2$ fixation[11], only enzymes of the WL pathway are present in LUCA (Supplementary Table 4). LUCA's WL enzymes are replete with FeS and FeNiS centres[21], indicating transition-metal requirements and also requiring organic cofactors: flavin, $F_{420}$, methanofuran, two pterins (the molybdenum cofactor MoCo and tetrahydromethanopterin) and corrins (Supplementary Table 4 and Supplementary Fig. 2). Microbes that use the WL pathway obtain their electrons from hydrogen[11], hydrogenases also being present among LUCA's genes. LUCA accessed nitrogen via nitrogenase and via glutamine synthetase. The WL pathway, nitrogenase and hydrogenases are also very oxygen-sensitive. LUCA was an anaerobic autotroph that could live from the gases $H_2$, $CO_2$ and $N_2$.

Several cofactor biosynthesis pathways trace to LUCA, including those for pterins, MoCo, cobalamin, siroheme, thiamine pyrophosphate, coenzyme M and $F_{420}$ (Supplementary Table 5). Many of these enzymes are *S*-adenosyl methionine (SAM)-dependent. A number of LUCA's SAM-dependent enzymes are radical SAM enzymes (Supplementary Table 1), an ancient class of oxygen-sensitive proteins harbouring FeS centres that initiate radical-dependent methylations and a wide spectrum of radical reaction mechanisms[22]. Radical SAM reactions point to a prevalence of one-electron reactions in LUCA's central metabolism, as does an abundance of flavoproteins (Supplementary Table 1), in addition to a prominent role for methyl groups.

FeS clusters, long viewed as relics of ancient metabolism[23,24], are the second most common cofactor/prosthetic group in LUCA's proteins behind ATP (Supplementary Table 1). The abundance of transition metals and FeS as well as FeNiS clusters in LUCA's enzymes indicates that it inhabited an environment rich in these metals. These features of LUCA's environment, in addition to thermophily and $H_2$, clearly point to a hydrothermal setting[12,13,17]. Selenoproteins, required in glutathione and thioredoxin synthesis and for some of LUCA's RNA modifications, are present, as is selenophosphate synthase (Supplementary Table 2). Like FeS centres, selenium in amino acids and nucleosides is thought to be an ancient trait[25]. Sulfur was involved in ancient metabolism[26] and LUCA was capable of S utilization, as indicated by siroheme, which is specific to redox reactions involving environmental S. Enzymes for sugar metabolism mainly encompass glycosylases, hydrolases and nonoxidative sugar metabolism, possibly reflecting primitive cell wall synthesis.

**LUCA's genes point to acetogenic and methanogenic roots.** The 355 trees also harbour phylogenetic information about LUCA's descendants, because archaea and bacteria are reciprocally rooted. Clostridia were the most frequently basal-branching bacteria, while methanogens were the most frequently basal-branching archaea (Supplementary Table 6). Clostridia and methanogens use the WL pathway[11]; they are abundant among microbial communities that inhabit the Earth's crust today[27,28], they harbour species that can live from methyl groups[6,27,28] and—like LUCA (Fig. 3)—they depend on $H_2$.

Today, environmental $H_2$ has two main sources: geological processes and $H_2$-producing fermentations. When LUCA existed, biological $H_2$ production did not exist, because primordial organics delivered from space are non-fermentable substrates[29]. For LUCA, that leaves only geological sources of $H_2$. The main geological source of $H_2$, both today and on the early Earth, is serpentinization, a process in which $Fe^{2+}$ in the crust reduces water circulating through hydrothermal systems to produce $H_2$ at high activities in hydrothermal effluent[30] of up to 26 mmol $kg^{-1}$. Yet, as well as $H_2$, methane[31,32] and other reduced C1 compounds[30,33] are synthesized abiotically in hydrothermal systems today.

**Hydrothermal vents, methyl groups, and nucleoside modifications.** LUCA's genes for RNA nucleoside modification (Supplementary Table 4) indicate that it performed chemical modification of nucleosides in both tRNA and rRNA[34]. Four of LUCA's nucleoside modifications are methylations requiring SAM (Supplementary Table 4). In the modern code, several base modifications are even strictly required for codon–anticodon interactions at the wobble position[35] (Supplementary Fig. 3). Consistent with the recurrent role of methyl groups in LUCA's biology, by far the most common tRNA and rRNA nucleoside modifications that are conserved across the archaeal bacterial divide[36] are methylations (Fig. 4a), although thiomethylations and incorporation of sulfur and selenium are observed.

That LUCA's genetic code involved modified bases in tRNA–mRNA–rRNA interactions attributes antiquity and functional



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

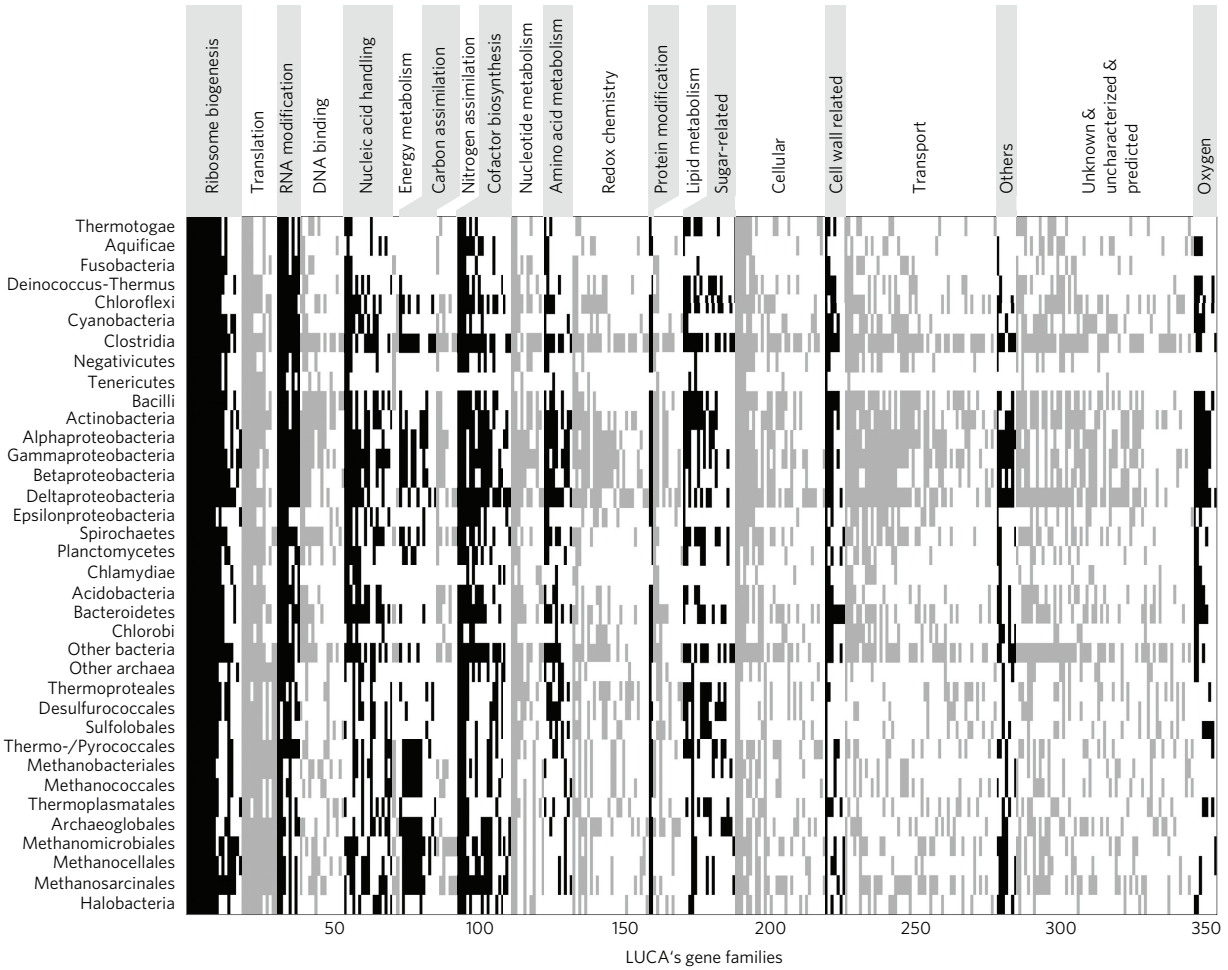


**Figure 2 | Taxonomic distribution of LUCA's genes grouped by functional categories.** The 355 clusters that fulfil the two criteria in Fig. 1 are grouped into 21 different functional categories (top). For each protein family (columns), ticks indicate the presence (black or grey) or absence (white) of the gene in the corresponding group shown on the left (rows). Within each functional category, clusters were sorted by the total sum of genes present (*x* axis). The distribution of the functional categories represented in the 355 LUCA gene families is significantly different from that represented in the data set as a whole (Supplementary Table 3).

significance to methylated bases in the evolution of the ribosome and the genetic code. It also forges links between the genetic code (Fig. 4a), primitive carbon and energy metabolism (Fig. 4b,c) and hydrothermal environments. How so? At modern hydrothermal vents, reduced C1 intermediates are formed during the abiotic synthesis of methane[30,33]. The intermediates can accumulate in some modern hydrothermal systems[33] and the underlying reactions can be simulated in the laboratory[37–39]. These reactions occur because under the reducing conditions of hydrothermal vents, the equilibrium in the reaction of $H_2$ with $CO_2$ lies on the side of reduced carbon compounds[40,41]. That is the reason why methanogens and clostridial acetogens, both of which figure prominently in autotrophic theories for life's origin[17,19], can grow by harnessing energy from the exergonic reactions of hydrogenotrophic methane and acetate synthesis[14,15]. The genes in LUCA's list (Supplementary Table 2) and the basal lineages among the 355 reciprocally rooted trees (Supplementary Table 6) indicate that LUCA lived in an environment where the geochemical synthesis of methane from $H_2$ and $CO_2$ was taking place, hence where chemically accessible reduced C1 intermediates existed.

Where LUCA arose, the genetic code arose. Either the chemical modifications of RNA nucleosides were absent in LUCA and were introduced later in evolution by some kind of adaptation, as some have suggested[42], or they are ancient[43]. Conservation of nucleoside modifications across the archaeal bacterial divide (Fig. 4a) indicate the latter. The enzymes that introduce nucleoside methylations are typically SAM enzymes, including members of the radical SAM family, which harbour an FeS cluster that initiates a radical in the reaction mechanism[22] and which are currently thought to be among the most ancient enzymes in metabolism[22]. Both the presence in LUCA's genome (Supplementary Table 4) of several SAM enzymes involved in nucleoside modifications and the presence of the nucleosides themselves (Fig. 4a), sometimes even at conserved positions in tRNA (Supplementary Fig. 3), indicate that these nucleoside methylations were present in LUCA's code, reflecting the code's ancestral state.

In methanogens and acetogenic clostridia, which the trees identified as the closest relatives of LUCA (Supplementary Table 6), methyl groups are central to growth, comprising the very core of carbon and energy metabolism. As shown in Fig. 4b, the methyl group generated by the WL pathway in the energy metabolism of methanogens is transferred from a nitrogen atom in tetrahydromethanopterin to a Co(I) atom in a corrin cofactor of the methyltransferase (MtrA-H) complex, possibly bound by MtrE[44], then subsequently transferred to the thiol sulfur atom of coenzyme M before being transferred to hydride at the methyl–CoM reductase



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

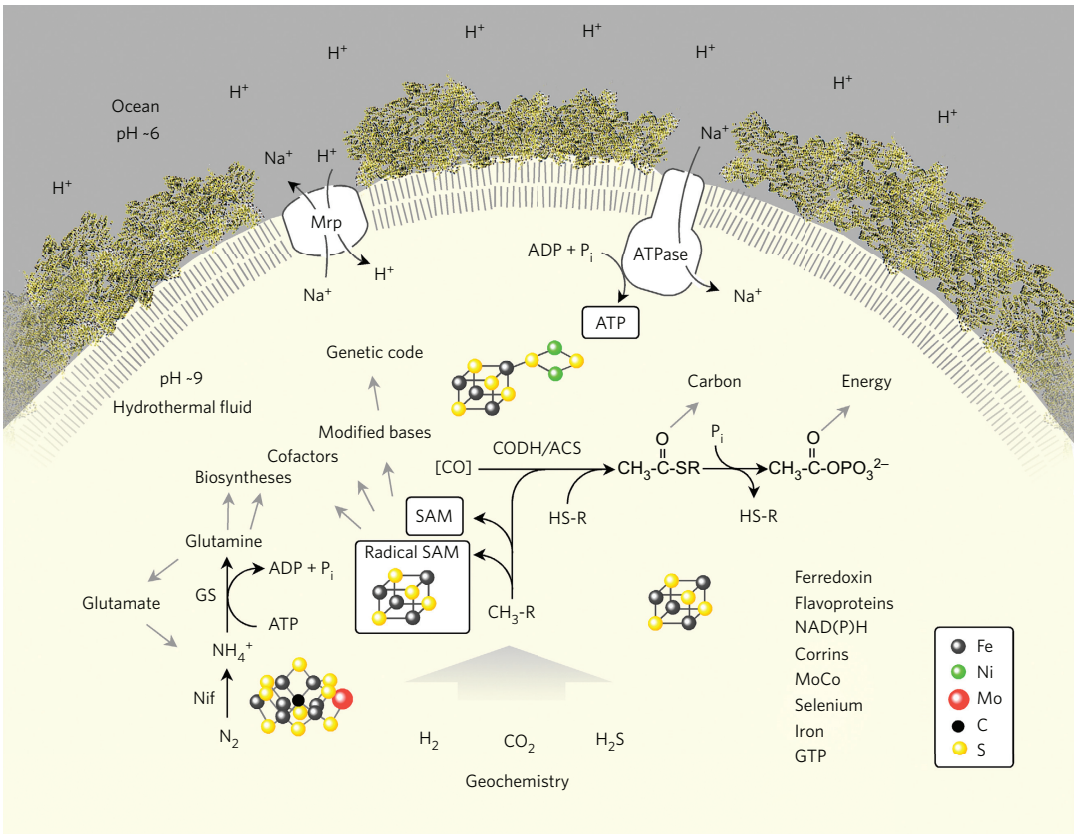


**Figure 3 | LUCA reconstructed from genome data.** Summary of the main interactions of LUCA with its environment, a vent-like geochemical setting[12,13,17,19], as inferred from genome data (Supplementary Table 2). Abbreviations: CODH/ACS, carbon monoxide dehydrogenase/acetyl CoA-synthase; Nif, nitrogenase; GS, glutamine synthetase; Mrp, MrP type $Na^+/H^+$ antiporter; $CH_3$-R, methyl groups; HS-R, organic thiols. The components listed on the lower right are present in LUCA, in addition to the cofactors listed in Supplementary Table 1. In modern CODH/ACS complexes, CO is generated from $CO_2$ and reduced ferredoxin[21]. The figure does not make a statement regarding the source of CO in primordial metabolism (uncatalysed via the gas water shift reaction or catalysed via transition metals), symbolized by [CO]. A $Na^+/H^+$ antiporter could transduce a geochemical pH gradient (indicated on the left) inherent in alkaline hydrothermal vents[13,17] into a more stable $Na^+$ gradient to feed a primordial Na-dependent ATP synthase[19]. LUCA undisputably possessed genes, because it had a genetic code; the question of which genes it possessed has hitherto been more difficult to address. The transition metal catalysts at the nitrogenase active site and the CODH/ACS active site as well as a 4Fe–4S cluster as in ferredoxin are indicated.

reaction. Energy conservation via $Na^+$ pumping occurs during the N-to-Co(I)-to-S transfer sequences of the MtrA-H complex[14,44]. This is an unusual coupling reaction in that electrons are not transferred; a methyl group is instead transferred, from a N atom to a S atom[14,44]. As shown in Fig. 4c, the methyl transfer chain of acetogens is a bit longer. It starts with a nitrogen-bound methyl moiety in tetrahydrofolate, which is transferred by AscE to a Co(I) atom in a corrin cofactor of the corrinoid FeS protein[45] and onto a Ni atom in the FeNiS cluster of acetyl-CoA synthase in an unusual metal-to-metal methyl transferase reaction[45]. Carbonyl insertion[21] generates a Ni-bound acetyl group that is removed via thiolysis to generate the thioester, which can either be used for carbon assimilation or for energy conservation as acyl phosphate and ATP[11,15]. In methanogens, carbon metabolism follows the same path to the thioester[11,14].

These methyl transfer reactions suggest that the environment where primordial carbon and energy metabolism arose was rich in methyl groups, S and transition metals. The conserved SAM-dependent methylations and S substitutions in modified nucleosides that allow tRNA anticodons to decode mRNA into protein carry the same chemical imprint (Fig. 4a and Supplementary Fig. 3), uncovering a hitherto underappreciated antiquity and significance of methyl groups at the core of biological chemistry. Methyl groups provide previously unrecognized links between carbon and energy metabolism in anaerobic autotrophs (Fig. 4b,c), tRNA–mRNA–rRNA interactions in the genetic code (Fig. 4a and Supplementary Fig. 3) and spontaneous chemistry at hydrothermal vents[30–33].

**Spelling out caveats and allowing for some LGT.** No approach to the study of early evolution is consummate; there are always caveats. Using our strict phylogenetic criterion, 355 protein families that are present in at least two higher taxa per domain and that preserve interdomain monophyly were identified. Universally distributed genes can be subject to transdomain LGT, yielding false negatives and underestimates of LUCA's gene content, while multiple LGT events might mimic vertical inheritance for some clusters, yielding false positives, or overestimates of LUCA's gene content. As an example of the latter, $O_2$-dependent enzymes should generally be absent from the list, because in LUCA's day, $O_2$ did not exist in physiologically relevant amounts[2]. LUCA's list does, however, contain five enzymes that use $O_2$ as a substrate and three that detoxify $O_2$ (Supplementary Table 2), functions that cannot be germane to LUCA, hence resulting from multiple transfers that phylogenetically emulate vertical intradomain inheritance. Given the massive influence that $O_2$ had on the origin and spread of new genes during evolution[46], finding $O_2$-dependent reactions at a frequency of 2.3% (8/355 proteins) suggests that the list of 355 genes harbours comparatively few multiple transfer cases. At the same time, ecological specialization to oxic niches will induce



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

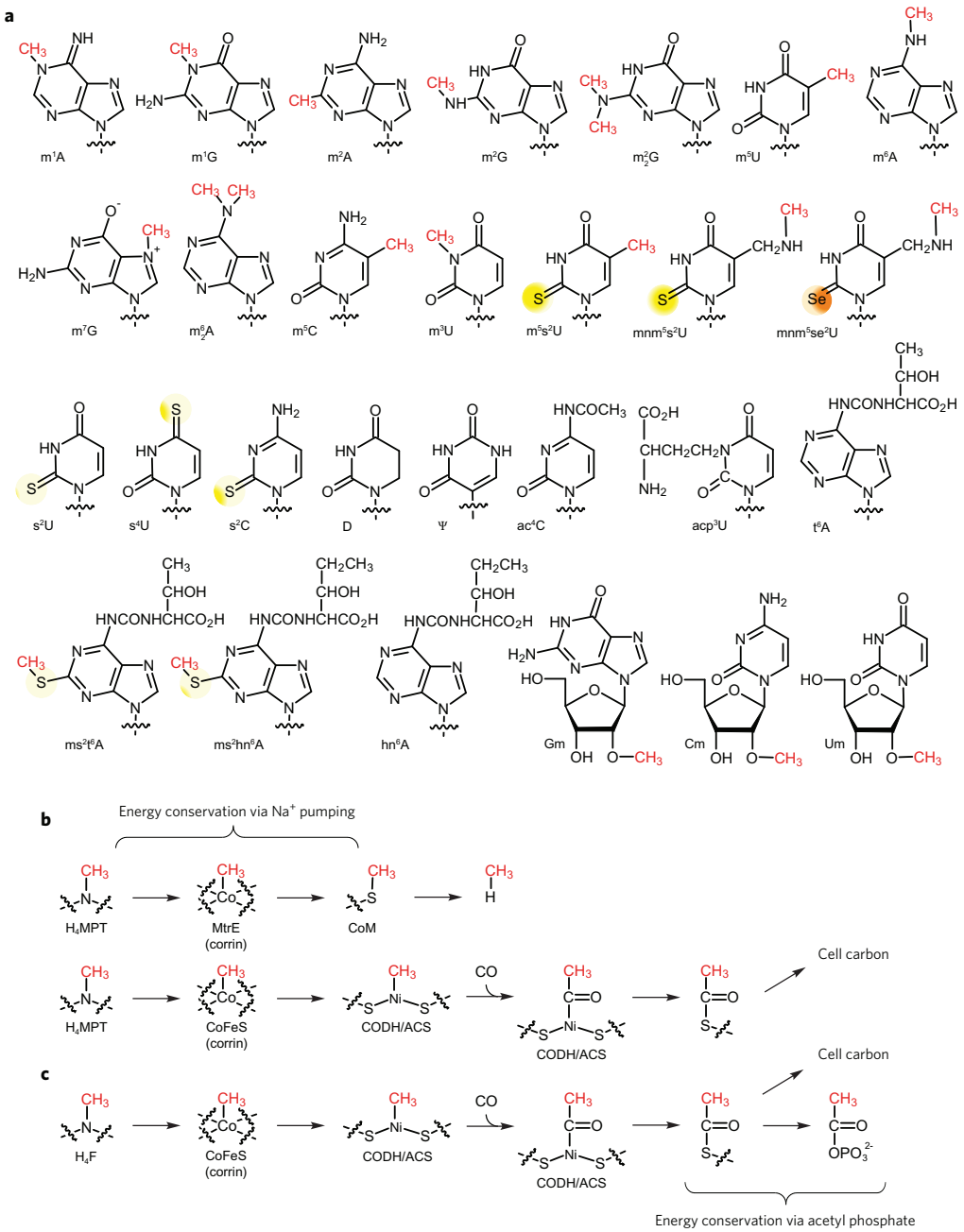


**Figure 4 | Methyl groups in conserved modified nucleosides and in anaerobic autotroph metabolism. a**, Structures of modified nucleosides found in tRNA and rRNA of archaea and bacteria[36]. Sulfur and selenium are highlighted in yellow and orange, respectively, while methyl groups are indicated in red. Many of these methylations are performed by SAM-dependent enzymes, which are mainly involved in nucleoside modifications and cofactor biosyntheses (Supplementary Table 7), suggesting a very central role of SAM in early metabolism and at the origin of tRNA–mRNA–rRNA interactions. **b**, In hydrogenotrophic methanogens, energy conservation occurs via $Na^+$ pumping during the N-to-Co(I)-to-S methyl transfer at the MtrA-H methyltransferase complex[14,44], while carbon assimilation[6] involves methyl transfer from $H_4$MPT to the corrinoid iron-sulfur protein (CoFeS)[45] and CODH/ACS[21], which reduces $CO_2$ to CO and catalyses the synthesis of acetyl-CoA. **c**, In hydrogenotrophic acetogens[15], the methyl group is transferred to CoFeS and CODH/ACS, where carbonyl insertion[21] and thiolysis to generate acetyl-CoA for carbon assimilation or for energy conservation[11] occur. Abbreviations: $m^1A$, 1-methyladenosine; $m^2A$, 2-methyladenosine; $m^6A$, $N^6$-methyladenosine; $t^6A$, $N^6$-threonylcarbamoyladenosine; $ms^2t^6A$, 2-methylthio-$N^6$-threonylcarbamoyladenosine; $m^5C$, 5-methylcytidine; Cm, 2′-O-methylcytidine; $ac^4C$, $N^4$-acetylcytidine; $m^1G$, 1-methylguanosine; $m^2G$, $N^2$-methylguanosine; $m^7G$, 7-methylguanosine; Gm, 2′-O-methylguanosine; $m^2_2G$, $N^2$,$N^2$-dimethylguanosine; ψ, pseudouridine; D, dihydrouridine; $m^5U$, ribosylthymine; Um, 2′-O-methyluridine; $s^2U$, 2-thiouridine; $s^4U$, 4-thiouridine; $m^5s^2U$, 5-methyl-2-thiouridine; $acp^3U$, 3-(3-amino-3-carboxyproyl)uridine; $mnm^5s^2U$, 5-methylaminomethyl-2-thiouridine; $mnm^5se^2U$, 5-methylaminomethyl-2-selenouridine; $s^2C$, 2-thiocytidine; $hn^6A$, $N^6$-hydroxynorvalylcarbamoyladenosine; $ms^2hn^6A$, 2-methylthio-$N^6$-hydroxynorvalyl carbamoyladenosine; $m^3U$, 3-methyluridine; $m^6_2A$, $N^6$,$N^6$-dimethyladenosine; $H_4$MPT, tetrahydromethanopterin; CoM, coenzyme M; CoFeS, corrinoid iron-sulfur protein; CODH/ACS, carbon monoxide dehydrogenase/acetyl-CoA synthase; $H_4$F, tetrahydrofolate; MtrE, a component of the MtrA-H methytransferase complex[44].



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

massive loss of anaerobe-specific genes in many lineages, leading to an underestimation of LUCA's gene content.

In addition, LUCA's gene list reveals only nine nucleotide biosynthesis and five amino acid biosynthesis proteins (Supplementary Table 2). The paucity of enzymes for essential amino acid, nucleoside and cofactor biosyntheses is most easily attributed to three factors: (1) the missing genes in question have been subject to interdomain LGT; (2) the genes are not well conserved at the sequence level, such that the bacterial and archaeal homologues do not fall into the same cluster; (3) LUCA had not yet evolved the genes in question prior to the bacterial–archaeal split, the pathway products for LUCA being provided by primordial geochemistry instead.

Low sequence conservation for proteins that were present in LUCA can also yield underestimates of LUCA's gene content. Because our criteria for presence in LUCA (domain monophyly) involve trees, the sequences need to be sufficiently well conserved to permit multiple sequence alignments and ML phylogenies, so a clustering threshold of 25% global identity was used. Yet genes that were present in LUCA that were not subject to transdomain LGT, but are not well conserved, can still fall into separate domain specific clusters. Relative to archaea, bacteria are overrepresented in the 355 families by a ratio of 134:1,847 in terms of genome sequences, for example in the far right of Supplementary Fig. 1. Despite our phylogenetic criteria, some LGTs might be among them. Enzymes for lipid metabolism in LUCA are scarce. The presence of a few enzymes involved in acyl-CoA metabolism might reflect multiple LGTs, as several archaea have acquired bacterial genes for fatty acid and aliphatic degradations[47].

Finally, one might ask what happens if we allow for a little bit of transdomain LGT? The minimum amount of transdomain LGT for which to allow would be reflected by a tree that fulfils our criteria of being present in two members each in two archaeal and two bacterial phyla (see Methods), but in addition, one bacterial sequence is misplaced within the archaea (or vice versa). If we allow for such cases representing one single transdomain LGT, then 124 new trees would be included (Supplementary Table 8), expanding our list to 479 members. If we allow for the next increment of LGT, namely that not one sequence but sequences from one archaeal phylum are misplaced within the bacteria (or vice versa), then 97 additional trees would be included (Supplementary Table 9), bringing the list to 576 proteins. The functional annotations in those expanded lists are very much in line with those reflected in the list of 355 summarized in Supplementary Tables 1 and 2.

## Discussion

Our findings clearly support the views that FeS and transition metals are relics of ancient metabolism[23,24], that life arose at hydrothermal vents[12,13], that spontaneous chemistry in the Earth's crust driven by rock–water interactions at disequilibrium thermodynamically underpinned life's origin[41,48] and that the founding lineages of the archaea and bacteria were $H_2$-dependent autotrophs that used $CO_2$ as their terminal acceptor in energy metabolism[17,19]. Spontaneous reactions involving C1 compounds and intermediate methyl groups in modern submarine hydrothermal systems[30–33] link observable geochemical processes with the earliest forms of carbon and energy metabolism in bacteria and archaea. In the same way that biochemists have long viewed FeS clusters as relics of ancient catalysis[23,24], methyl groups appear here as relics of primordial carbon and energy metabolism.

Although the paucity in LUCA of genes for amino acid and nucleoside biosyntheses could, in principle, be attributable to post-LUCA LGT, we note that there is no viable alternative to the view that LUCA, regardless of how envisaged, ultimately arose from components that were synthesized abiotically via spontaneous, exergonic syntheses somewhere during the history of early Earth[48]. Prior to the origin of genes, proteins and the code, LUCA's origin was hence dependent on spontaneous organic syntheses, which are thermodynamically favourable under the high $H_2$ activities of submarine hydrothermal vents[40,41], and which still occur today in some geochemical environments[30–33]. The notion that early replicating systems tapped environmental supplies of biologically relevant compounds provided by spontaneous (exergonic) chemical reactions might seem to be a very radical proposition in the present Article, but it is inherent, often implicitly, to all theories for the prebiotic origin of replicating systems and life.

Genome data depict LUCA as a strictly anaerobic, $H_2$-dependent thermophilic, diazotrophic autotroph with a WL pathway and that lived in a hydrothermal vent setting. These are attributes of acetogenic clostridia and methanogens, lineages that branch deeply in trees of LUCA's genes and that occupy the Earth's crust today. Methyl groups were central to carbon and energy metabolism in LUCA, and they also persist to the present as chemical relicts in tRNA–mRNA interactions at the ribosome, suggesting that LUCA not only lived in a hydrothermal vent setting rich in $H_2$, $CO_2$, transition metals, sulfur and reactive C1 species of geochemical origin, but that LUCA's genetic code arose there as well. The data provide evidence in favour of autotrophic origins[11] over heterotrophic origins[49] and converge with independent geochemical evidence[30–33], favouring theories that posit a single hydrothermal environment rich in $H_2$ and transition metals for LUCA's origin[11–13,16,17,19,48] over theories that entail many different kinds of chemical environments[50] catalysing one reaction each.

## Methods

**Sequence clustering and gene phyletic pattern reconstruction.** Protein sequences of 1,981 complete prokaryotic genomes were downloaded from the NCBI RefSeq[51] database (version June 2012). These genomes were grouped into 13 archaeal and 23 bacterial groups corresponding to NCBI taxonomic orders, phylum or class, respectively, as in ref. 9. Markov chain clustering[52] (MCL) of sequences was performed as previously described[9], with the reciprocal best BLAST[53] (v. 2.2.28) hit (rBBH) procedure and an *E*-value threshold of $\leq 10^{-10}$. rBBH-pairs with global amino acid identity not smaller than 25%, calculated with needle[54] from EMBOSS 6.6.0.0, were clustered using MCL. Proteins with no significant homologues were classified as singletons and discarded from further analysis. Protein families with archaeal and bacterial sequences were considered further.

**Multiple sequence alignment and reconstruction of phylogenetic trees.** Sequences in each of the MCL clusters were aligned using MAFFT[55] version 7.130 with the options *–localpair*, *–maxiterate = 1000* and *–anysymbol*. The heads-or-tails method[56,57] was used to compare alignment reliability for different sets of clusters with an inbuilt program. ML trees were reconstructed using RAxML[58] v7.8.6 under the PROTCATWAG model, with special amino acid characters U, O and J converted into C, K and X. These trees were rerooted between archaea and bacteria and parsed for monophyly of sets of sequences using the programs *nw_reroot* and *nw_clade* with the *-m* option from Newick Utilities[59] version 1.6. Single group paraphyly was identified as those trees where a single archaeal or bacterial group was placed within the other domain.

**LUCA gene identification.** Phyletic patterns for 1,981 genomes were analysed and a gene family was only considered as present in LUCA if there were at least two archaeal groups and two bacterial groups, each of which had a minimum of two members, respectively, present in the given family and if archaea and bacteria were monophyletic in the corresponding phylogenetic tree. Exceptions were made for the four under-represented groups ('Methanocellales', 'Thermoplasmatales', 'Archaeoglobales' and 'other archaea'), where the presence of only one sequence instead of a minimum of two was counted as present.

**Functional annotation and cofactor determination.** The protein families were annotated using COG[60] and KEGG[61] functional categories. COG identifiers (COG IDs) of all sequences from 355 candidate LUCA clusters were extracted from the NCBI RefSeq database (June 2012). A particular COG ID was assigned to a cluster if it was assigned to more sequences in that cluster than any other COG ID. Each COG ID was then mapped to the COG functional categories. If a COG ID mapped to more than one category, the category R (general function prediction only) was assigned. The same procedure was repeated to annotate families using the KEGG database. In addition, for each protein family, the most frequent PFAM-A domain annotation (version 28.0, June 2015) was obtained by using the HMM approach as available at PFAM[62]. The identification of protein cofactors and catalytic centres present in the families as well as their organization into categories was performed through case by



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

case inspection of each of the 355 families. The functional categories listed in Supplementary Tables 1, 2, 4 and 5 are designed to reflect microbial physiology (for example, the category redox), so they do not correspond 1:1 to COG or KEGG functional categories; however, the COG and KEGG categories for each protein are given in Supplementary Table 2. For a test of independence (see section 'Test of independence'), COG categories were used. Cofactors occurring only in one protein family are not shown in Supplementary Table 1. These include thiamine pyrophosphate, pyridoxal phosphate, biotin, methyltetrahydrofolate, menaquinone, iron and rubredoxin. Copper (two occurrences) is also not shown because it is only present as a metal centre in oxygen-related protein families. For protein families corresponding to subunits of a protein complex, a subunit was scored as presence of the complex, such that cofactors of the complex were counted even if not all subunits were present in our list. For complexes scored as present, cofactors were counted only once. Ferredoxin, flavodoxin and methanophenazine were counted as ferredoxin because proteins annotated as coenzyme $F_{420}$-reducing hydrogenase can correspond to protein families that bind either ferredoxin or methanophenazine, while the electron carrier for NifH can be either ferredoxin or flavodoxin. Thioredoxin was not counted as a cofactor.

**Presence of selenoproteins.** All amino acid sequences present in the 355 clusters were searched for the presence of the one letter code amino acid 'U' that represents selenocysteine. Selenium was considered present in a protein family when selenocysteine was found in at least one protein sequence of the cluster.

**Deeply branching archaeal and bacterial lineages.** Deeply branching groups in each domain were identified based on two different factors: (1) the smallest split that contains both archaea and bacteria, or (2) sequences connecting both archaea and bacteria with shortest evolutionary distance in the tree. The smallest split that contains both archaea and bacteria was obtained by parsing the bipartitions of the tree, and two different cases were considered: (1) in the $Tree_{Pure}$ method, intradomain basal branches containing sequences from only one phylum/group were counted; (2) in the $Tree_{Mixed}$ method, the intradomain basal branches containing sequences belonging to organisms from more than one phylum/group were counted. To identify sequences connecting archaea and bacteria at the shortest evolutionary distance, trees were translated into a distance matrix of branch lengths using the *nw_distance* program from Newick Utilities[59] (version 1.6). The sequence pair with shortest distance connecting archaea and bacteria was identified from this distance matrix ($Dist_{Root}$ method).

**Ribosome, tRNA and nucleoside modifications.** Sequences of prokaryotic tRNA containing modifications were downloaded from the Modomics[63] and tRNAdb[64] databases (December 2014). Bacterial and archaeal tRNA modifications were mapped into the tRNA structure and common modifications occurring at the same tRNA position in both prokaryotic domains were identified. Supplementary Fig. 3a was prepared using VMD[65] version 1.9.2. All chemical structures were drawn in ACD/ChemSketch (2015 release, Advanced Chemistry Development; www.acdlabs.com).

**Test of independence.** To determine whether functional annotations and taxonomical groups were distributed non-randomly in the 355 candidate LUCA clusters, a $\chi^2$ test of independence was performed. All 11,093 clusters that contained homologues from both archaea and bacteria were used to calculate the expected distribution. COG categories were separated into five groups: information, metabolism, cellular, poorly characterized and not declared (including clusters that were not annotated by the COG database). There were 36 taxonomical groups used for the lineage distribution: 13 archaeal groups (Archaeoglobales, Desulfurococcales, Halobacteria, Methanobacteriales, Methanocellales, Methanococcales, Methanomicrobiales, Methanosarcinales, Sulfolobales, Thermococcales/Pyrococcales, Thermoplasmatales, Thermoproteales, other archaea) and 23 bacterial groups (Acidobacteria, Actinobacteria, Alphaproteobacteria, Aquificae, Bacilli, Bacteroidetes, Betaproteobacteria, Chlamydiae, Chlorobi, Chloroflexi, Clostridia, Cyanobacteria, Deinococcus-Thermus, Deltaproteobacteria, Epsilonproteobacteria, Fusobacteria, Gammaproteobacteria, Negativicutes, Planctomycetes, Spirochaetes, Tenericutes, Thermotogae, other bacteria). The distribution of functional annotations and taxonomical groups of the 355 candidate LUCA clusters were compared with those expected (degrees of freedom = 4 and degrees of freedom = 35, respectively). *P* values were calculated using MATLAB R2015a and its function chi2cdf.

Received 19 April 2016; accepted 21 June 2016; published 25 July 2016

## References


1. Fox, G. E. *et al.* The phylogeny of prokaryotes. *Science* **209,** 457–463 (1980).
2. Arndt, N. & Nisbet, E. Processes on the young Earth and the habitats of early life. *Annu. Rev. Earth Planet Sci.* **40,** 521–549 (2012).
3. Woese, C. The universal ancestor. *Proc. Natl Acad. Sci. USA* **95,** 6854–6859 (1998).
4. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1,** 127–136 (2003).
5. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–236 (2013).
6. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl Acad. Sci. USA* **112,** 6670–6675 (2015).
7. Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Res. Microbiol.* **157,** 57–68 (2006).
8. Kannan, L., Li, H., Rubinstein, B. & Mushegian, A. Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biol. Direct.* **8,** 32 (2013).
9. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517,** 77–80 (2015).
10. Say, R. F. & Fuchs, G. Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* **464,** 1077–1081 (2010).
11. Fuchs, G. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65,** 631–658 (2011).
12. Baross, J. A. & Hoffman, S. E. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins Life Evol. B* **15,** 327–345 (1985).
13. Russell, M. J. & Hall, A. J. The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J. Geol. Soc. Lond.* **154,** 377–402 (1997).
14. Buckel, W. & Thauer, R. K. Energy conservation via electron bifurcating ferredoxin reduction and proton/$Na^+$ translocating ferredoxin oxidation. *Biochim. Biophys. Acta* **1827,** 94–113 (2013).
15. Schuchmann, K. & Müller, V. Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria. *Nature Rev. Microbiol.* **12,** 809–821 (2014).
16. Ferry, J. G. & House, C. H. The step-wise evolution of early life driven by energy conservation. *Mol. Biol. Evol.* **23,** 1286–1292 (2006).
17. Martin, W. & Russell, M. J. On the origin of biochemistry at an alkaline hydrothermal vent. *Phil. Trans. R. Soc. Lond. B* **362,** 1887–1925 (2007).
18. Mulkidjanian, A. Y., Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Evolutionary primacy of sodium bioenergetics. *Biol. Direct.* **3,** 13 (2008).
19. Lane, N. & Martin, W. F. The origin of membrane bioenergetics. *Cell* **151,** 1406–1416 (2012).
20. Déclais, A. C., Marsault, J., Confalonieri, F., La Tour de, C. B. & Duguet, M. Reverse gyrase, the two domains intimately cooperate to promote positive supercoiling. *J. Biol. Chem.* **275,** 19498–19504 (2000).
21. Ragsdale, S. W. Nickel-based enzyme systems. *J. Biol. Chem.* **284,** 18571–18575 (2009).
22. Broderick, J. B., Duffus, B. R., Duschene, K. S. & Shepard, E. M. Radical *S*-adenosylmethionine enzymes. *Chem. Rev.* **114,** 4229–4317 (2014).
23. Eck, R. V. & Dayhoff, M. O. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152,** 363–366 (1966).
24. Hall, D. O., Cammack, R. & Rao, K. K. Role of ferredoxins in the origin of life and biological evolution. *Nature* **233,** 136–138 (1971).
25. Böck, A., Forchhammer, K., Heider, J. & Baron, C. Selenoprotein synthesis: an expansion of the genetic code. *Trends Biochem. Sci.* **16,** 463–467 (1991).
26. Liu, Y. C., Beer, L. L. & Whitman, W. B. Methanogens: a window into ancient sulfur metabolism. *Trends Microbiol.* **20,** 251–258 (2012).
27. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350,** 434–438 (2015).
28. Lever, M. A. Acetogenesis in the energy-starved deep biosphere—a paradox? *Front. Microbiol.* **2,** 284 (2012).
29. Schönheit, P., Buckel, W. & Martin, W. F. On the origin of heterotrophy. *Trends Microbiol.* **24,** 12–25 (2016).
30. Schrenk, M. O., Brazelton, W. J. & Lang, S. Q. Serpentinization, carbon, and deep life. *Rev. Mineral. Geochem.* **75,** 575–606 (2013).
31. Etiope, G. & Schoell, M. Abiotic gas: atypical, but not rare. *Elements* **10,** 291–296 (2014).
32. Proskurowski, G. *et al.* Abiogenic hydrocarbon production at Lost City hydrothermal field. *Science* **319,** 604–607 (2008).
33. McDermott, J. M., Seewald, J. S., German, C. R. & Sylva, S. P. Pathways for abiotic organic synthesis at submarine hydrothermal fields. *Proc. Natl Acad. Sci. USA* **112,** 7668–7672 (2015).
34. Chow, C. S., Lamichhane, T. N. & Mahto, S. K. Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. *ACS Chem. Biol.* **2,** 610–619 (2007).
35. Agris, P. F., Vendeix, F. A. P. & Graham, W. D. tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol.* **366,** 1–13 (2007).
36. Grosjean, H., Gupta, R., & Maxwell, E. S. in *Archaea: New Models for Prokaryotic Biology* (ed. Blum, P.) 171–196 (Caister Academic Press, 2008).




© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036

37. Seewald, J. S., Tolotov, M. Y. & McCollom, T. Experimental investigation of single carbon compounds under hydrothermal conditions. *Geochim. Cosmochim. Acta* **70,** 446–460 (2006).
38. He, C., Tian, G., Liu, Z. & Feng, S. A mild hydrothermal route to fix carbon dioxide to simple carboxylic acids. *Org. Lett.* **12,** 649–651 (2010).
39. Horita, J. & Berndt, M. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285,** 1055–1057 (1999).
40. Amend, J. P. & Shock, E. L. Energetics of amino acid synthesis in hydrothermal ecosystems. *Science* **281,** 1659–1662 (1998).
41. Amend, J. P., LaRowe, D. E., McCollom, T. M. & Shock, E. L. The energetics of organic synthesis inside and outside the cell. *Phil. Trans. R. Soc. Lond. B* **368,** 20120255 (2013).
42. Yokoyama, S., Watanabe, K. & Miyazawa, T. Dynamic structures and functions of transfer ribonucleic acids from extreme thermophiles. *Adv. Biophys.* **23,** 115–147 (1987).
43. Helm, M. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Res.* **34,** 721–733 (2006).
44. Gottschalk, G. & Thauer, R. K. The $Na^+$-translocating methyltransferase complex from methanogenic archaea. *Biochim. Biophys. Acta* **1505,** 28–36 (2001).
45. Svetlitchnaia, T., Svetlitchnyi, V., Meyer, O. & Dobbek, H. Structural insights into methyltransfer reactions of a corrinoid iron–sulfur protein involved in acetyl-CoA synthesis. *Proc. Natl Acad. Sci. USA* **103,** 14331–14336 (2006).
46. Raymond, J. & Segre, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311,** 1764–1767 (2006).
47. Dibrova, D. V., Galperin, M. Y. & Mulkidjanian, A. Y. Phylogenomic reconstruction of archaeal fatty acid metabolism. *Environ. Microbiol.* **16,** 907–918 (2014).
48. Shock, E. L. & Boyd, E. S. Geomicrobiology and microbial geochemistry: principles of geobiochemistry. *Elements* **11,** 389–394 (2015).
49. Mansy, S. S. *et al.* Template-directed synthesis of a genetic polymer in a model protocell. *Nature* **454,** 122–125 (2008).
50. Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D. & Sutherland, J. D. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nature Chem.* **7,** 301–307 (2015).
51. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** D130–D135 (2011).
52. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An ancient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
53. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
54. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
56. Landan, G. & Graur, D. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24,** 1380–1383 (2007).
57. Landan, G. & Graur, D. Local reliability measures from sets of co-optimal multiple sequence alignments. *Pac. Symp. Biocomput.* **13,** 15–24 (2008).
58. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
59. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26,** 1669–1670 (2010).
60. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
61. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27,** 29–34 (1999).
62. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).
63. Machnicka, M. A. *et al.* MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res.* **41,** D262–D267 (2013).
64. Jühling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **37,** D159–D162 (2009).
65. Humphrey, W., Dalke, A. & Schulten, K. VMD—visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).



## Acknowledgements

The authors thank J. Baross and N. Lane for discussions. The authors acknowledge the Zentrum für Informations- und Medientechnologie (ZIM) of the Heinrich-Heine University for computational support and the European Research Council for funding (ERC AdG 666053 to W.F.M.).


## Author contributions

M.C.W., F.L.S., S.N., M.R. and S.N.-S. performed the bioinformatics analysis. F.L.S. and N.M. carried out the functional classification of the protein families. All authors analysed and discussed the results. W.F.M., F.L.S. and S.N.-S. designed the research. M.C.W., F.L.S., S.N., N.M., S.N.-S. and W.F.M. wrote the paper.

## Additional information

Supplementary information is available online. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to W.F.M.

## Competing interests

The authors declare no competing financial interests.



© 2016 Macmillan Publishers Limited. All rights reserved

Bayer EX1083
Bayer v. Corteva - IPR2023-01036